# The Application of Atheoretical Regression Trees to Problems in Time Series Analysis

A thesis submitted in partial fulfilment of the
requirements for the Degree of
Doctor of Philosophy in Mathematics

by William S. Rea
Department of Mathematics and Statistics
University of Canterbury

2008

# Contents

# Acknowledgments

I would like to thank the following people. My supervisors Dr Marco Reale, Associate Professor Jennifer Brown and Professor Les Oxley without whom this thesis would never have been completed. Sir Clive Granger and Professor Peter Robinson for some very useful discussions on long memory time series. Arek Ohanissian, Jeffrey Russell, and Ruey Tsay for providing us with the source code for their test. Marcel Scharth and Marcelo Medeiros for providing us with a copy of their realized volatility data. Alethea Rea for many interesting and helpful conversations, reading many drafts of various chapters and papers and for listening to many practice sessions of presentations. Dr Dominic Lee for useful discussions on the use of the bootstrap. Attendees at the 2006, 2007, and 2008 New Zealand Econometric Study Group Meetings, the 2006 and 2007 International Workshops on Statistical Modeling, the MODSIM 2007 conference, and the 2007 Workshop on Non-linear and Complex Systems Analysis for their positive and helpful feedback on the research presented at those conferences. Editors and anonymous referees from the following journals for their useful feedback on papers arising from this thesis; Journal of Computational Statistics and Data Analysis, Journal of Economic and Business Statistics and Econometric Reviews. The three examiners, Professor Aaron Smith, Dr Laimonis Kavalieris, and Dr Chris Price, whose comments improved the final version of the thesis. The Information and Communication Technology Services of the University of Canterbury for supporting me with study and conference leave to pursue this research. Finally I would like to acknowledge the constant support of my wife, Joy, without whom this thesis could not have been undertaken.

**Abstract**

This thesis applies Atheoretical Regression Trees (ART) to the problem of locating changes in mean in a time series where the number and location of those changes are unknown. We undertook an extensive simulation study into ART's performance on a range of time series. We found ART to be a useful addition to currently established structural break methodologies such as the CUSUM and that due to Bai and Perron. ART was found to be useful in the analysis of long time series which are not practical to analyze with the optimal procedure of Bai and Perron.

ART was applied to a long standing problem in the analysis of long memory time series. We propose two new methods based on ART for distinguishing between true long memory and spurious long memory due to structural breaks. These methods are fundamentally different from current tests and procedures intended to discriminate between the two sets of competing models. The methods were subjected to a simulation study and shown to be effective in discrimination between simple regime switching models and fractionally integrated processes.

We applied the new methods to 16 realized volatility series and concluded they were not fractionally integrated series. All 16 series had mean shifts, some of which could be identified with historical events.

We applied the new methods to a range of geophysical time series and concluded they were not fractional Gaussian noises. All of the series examined had mean shifts, some of which could be identified with known climatic changes.

We conclude that our new methods are a significant advance in model discrimination in long memory series.

# Chapter 1

# Introduction

This thesis deals with the application of Atheoretical Regression Trees (ART), a structural break detection and location method, to problems in time series analysis and model discrimination in long memory time series. This chapter reviews some of the relevant literature. The scope of the thesis is such that several diverse, but sometimes inter-related, areas of the literature are reviewed. These are as follows.

1. The literature on structural breaks, including ART.
2. The literature on long memory time series from a true long memory point of view.
3. The literature on estimators of the long memory parameters $H$ and $d$ (defined in Section 6.3 below) and related goodness-of-fit tests.
4. The literature which deals with the duality between breaks and long memory in time series and existing tests to discriminate between them.
5. The literature from marketing, geophysics, paleoclimatology, economics and finance relevant to each of the three case studies.

Each of these areas is large with the exception of the marketing literature relevant to the Crest toothpaste data. Thus it is necessary to be selective to keep the review to a reasonable size. We review the literature for point one in this chapter, points two,

three and four in Chapters (6), (7), and (8) respectively and point five in the relevant chapters of the case studies.

## 1.1 Structural Break Model and Tests

We define the break model as follows:

$$y_t = \mu_{y_t} + \epsilon_t \tag{1.1}$$

$$\mu_{y_t} = \sum_{i=1}^{p} I_{(t_{i-1}<t\leq t_i)}\mu_i \tag{1.2}$$

where $y_t$ is the observed value at time $t$, $\epsilon_t$ is an error term which may have serial correlation, may be non-normal but is assumed stationary, $\mu_{y_t}$ is the mean of the time series at $t$, $I_{t\in R}$ is an indicator variable which is 1 only if $t \in R$, and 0 otherwise, $t_i$, $i = 1, \ldots, p$, are the breakpoints, $t_0$ is defined to be time zero whereas the observations start at $t = 1$ and $\mu_i$ is the mean of the regime $i$. The regime $R$ is defined as the period between breakpoints $t_{i-1} < t \leq t_i$.

It is important to note that Equations (1.1) and (1.2) are just a way to represent a single time series as a sequence of different models (i.e. a series subjected to structural breaks). Equations (1.1) and (1.2) only deal with breaks in the mean and can be generalized to model breaks in other parameters. In real series a structural break signals a discontinuity in the data generating process (DGP) and so the parameters for each regime must be estimated from data within the regimes. We are considering this class of model because it has been used by many others, including when studying long memory processes.

Csörgo and Horvath (1997) produced an extensive monograph on the subject of change-point analysis. Since the publication of that book there have been two recent reviews of structural breaks, Hansen (2001), and the overview to the Journal of Econometrics, special issue on structural breaks, long memory and stock market volatility of Banerjee and Urga (2005).

There are a number of tests which have been developed to test for the presence of structural breaks, often in the context of regression. Some are as follows.

1. Quandt (1958) considered several tests which were based on the exhaustive calculation of the log-likelihood function of the data obeying two different regression regimes using all possible binary divisions of the data. For a test of hypothesis of two regimes against an alternative of one regime he considered two possible tests.

   (a) A large sample test under which he proposed the likelihood ratio followed a $\chi^2$-distribution.

   (b) A small sample test under which he proposed the likelihood ratio followed an F-distribution. A variant of this test omitted the middle portion of the data.

   Quandt (1960) reconsidered the problem and rejected the $\chi^2$-distribution in the large sample test. He replaced it with a table of empirical critical values. He also discarded the assumption of independent, identically distributed (*iid*) $N(0, \sigma^2)$ residuals in favour of a $t$-distribution. He proposed a new test based on standardized residuals described by an $F$-distribution.

   Ferreira (1975) considered the tests of Quandt in a Bayesian framework under the assumption that the number of regimes was known *a priori*. Ferreira provided formulas for calculating the posterior probability function or density for the breakpoint and all regime parameters.

2. Chow (1960) considered the problem of having estimated a linear regression from $p$ observations whether $m$ new observations belonged to the same regression. In the case $m < p$ two tests were considered.

   (a) The $m$ new data points were compared with their prediction interval from the previously estimated regression.

   (b) An $F$-ratio for testing expected values of the residuals against the actual values.

   He further extended his tests to two subsets of the $p$ observations.

   In the case $m \geq p$ he proposed an analysis of covariance test.

3. The CUSUM (CUmulative SUMmation) family, see Page (1954), Barnard (1959), Brown et al. (1975), Pettitt (1980), Sen (1982), Ploberger and Krämer (1992) and Sibbertsen (2000). The CUSUM test is described in Section (1.1.1) below.

4. Generalized Method of Moments (GMM). While versions of GMM have been used for many years, the general statement was developed by Hansen (1982). A good summary is contained in the textbook of Hamilton (1994, pp409–434). The main advantage of this method is that it does not require the full probability density to be specified, only certain moment conditions. This is also a weakness because often GMM will not make efficient use of all the information in a sample.

   Conceptually it is not difficult to understand that if there were one or more structural breaks present in the sample, the sample moments will change across the break(s). GMM locates candidate structural breaks by testing for changes in the estimated moments of a time series.

   The highly cited paper of Andrews (1993) presents Wald, Lagrange Multiplier and Likelihood ratio tests based on GMM. The recent paper of Gagliardini et al. (2005) presents new GMM tests for structural breaks and numerous references to literature on GMM tests.

5. Andrews and Ploberger (1994) presented exponentially weighted versions of the Wald, Lagrange Multiplier and Likelihood ratio tests to specifically take account of the fact that in structural break tests there are parameters present in the alternative hypothesis which are not present under the null. The location of the breakpoint(s) are an example. The existence of these "nuisance" parameters gives the tests non-standard distributions which must be derived separately from the case where the number and meaning of the parameters do not change under the different hypotheses.

6. Optimal tests of Bai and Perron (1998, 2003) based on the work of Fisher (1958), Bell and Roth (1969) and Guthery (1974). See Section (1.1.2) below for details.

7. Methods based on the Classification And Regression Tree (CART) methodology of Breiman et al. (1993). See Cappelli and Reale (2005), Rea et al. (2006), Cappelli et al. (2007), and Cappelli et al. (2008). These are discussed in Section (1.2) and following below.

In this thesis, in addition to regression trees, we also use the CUSUM test and the procedure of Bai and Perron (1998, 2003) which we now review.

### 1.1.1 The CUSUM Tests

Cumulative summation, or CUSUM, tests proposed by Page (1954) have their origin in industrial quality control as a simple graphical method of detecting small changes in process parameters. The original method involved taking a sample from the process, calculating a score, $x_t$, for the sample and plotting the cumulative sum of the scores

$$S_n = \sum_{k=1}^{n} x_k$$

on a chart. Curiously, the procedure was originally designed so that if the graph was tracking downwards the process was performing satisfactorily. Barnard (1959), among other contributions, debated the merits of triangular and parabolic shaped pieces of cardboard which could be placed over the charts by process operators to assess its current state.

Plotting either the ordinary least-squares residuals or their squares against time is not a sensitive indicator of small or gradual changes in regression parameters. Thus in a CUSUM test instead of plotting the individual residuals, $z_t$, the cumulative summation

$$Z_r = \frac{1}{\hat{\sigma}} \sum_{t=1}^{r} z_t; r = 1, \ldots, T \tag{1.3}$$

is plotted against time. The term $1/\hat{\sigma}$, where $\hat{\sigma}$ is the sample estimated standard deviation, standardizes the scale of the summation. The graph is then examined to see if there is a systematic departure of the $Z_r$ from the $E[Z_r] = 0$ line, where $E[\cdot]$ denotes expectation, as in Figure (1.1).

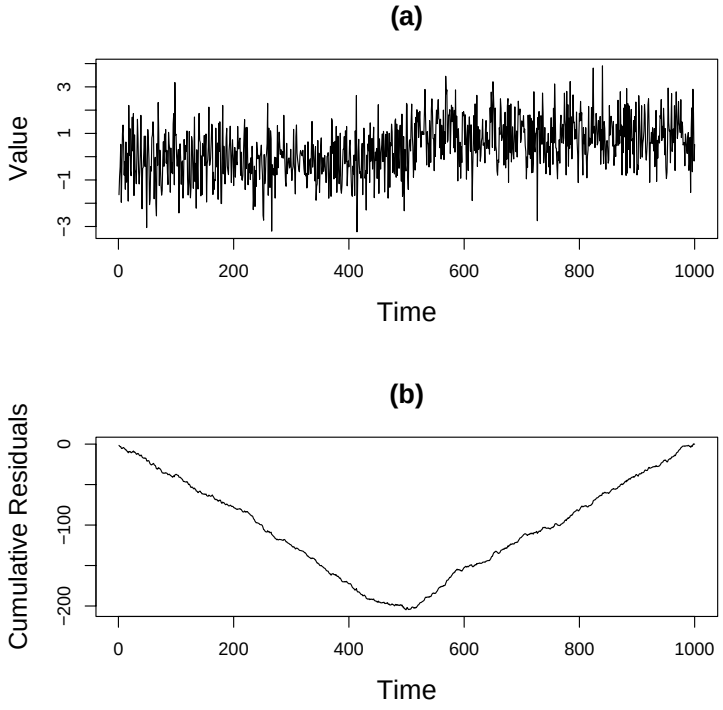


Figure 1.1: Panel (a) is a time series with a structural break at $T = 500$ and in panel (b) is its corresponding CUSUM plot.

Initially the test was largely graphical. Brown et al. (1975) developed formal statistical tests of significance by drawing on the theory of stochastic processes as covered in standard texts such as those of Cox and Miller (1965), Karlin and Taylor (1975) and Hoel et al. (1987). Under the null hypothesis the sum in Equation (1.3) is a Wiener process. The problem becomes one of calculating the probability the sample path of Equation (1.3) will cross a given line, hence determining lines of appropriate levels of statistical significance.

CUSUM tests have continued to be developed. Ploberger and Krämer (1992) extended the tests to use ordinary least squares (OLS) residuals. Manly and MacKenzie (2000) presented modifications to handle problems in environmental monitoring and increased the robustness of the tests to serial correlations. Sibbertsen (2000) examined

the CUSUM-M test in the presence of long-memory disturbances.

In this thesis they are used for comparison with ART in detecting structural breaks. Unless otherwise indicated we use the CUSUM tests as implemented in the `R` software of R Development Core Team (2005) contributed package `strucchange` of Zeileis et al. (2002).

### 1.1.2 The Bai and Perron Procedure

The Bai and Perron procedure (hereafter BP) for finding structural breaks in linear regressions was presented in two papers. Bai and Perron (1998) dealt with theoretical results. Bai and Perron (2003) dealt with the empirical implementation of the procedure. Their method, as was ART, was based on the Fisher (1958) algorithm to find the least squares partition of $T$ contiguous objects into $G$ groups. BP also drew on later work by Bell and Roth (1969) and Guthery (1974).

The Fisher algorithm required $O(K^2G)$ steps to partition $K$ elements into $G$ exhaustive and mutually exclusive groups. Bai and Perron were able to considerably improve on the computational efficiency of the algorithm in a number of ways.

Using BP's notation, for a sample size of $T$ the total number of possible segments was at most $T(T+1)/2$. However, in practice this limit was never reached.

BP imposed three conditions. They were:

1. The minimum distance between breaks was $h$.

2. For a series with $m$ breaks and $m+1$ regimes the largest segment must be sufficiently small to allow $m$ other segments before and/or after it.

3. A segment could not begin at dates 2 through $h$ as this would violate condition one. A similar restriction applied to where a regime could end.

Once these conditions were imposed the number of allowable segments had been considerably reduced. BP then calculated a somewhat sparsely populated upper triangular matrix of sum of squared residuals (SSR) for the allowable segments. An example of this upper triangular matrix is presented in Figure (1) of Bai and Perron

(2003). The full matrix would have been symmetric eliminating the need to calculate the lower triangle. (On a practical note BP is very computationally intensive. Most of the processing time is spent calculating the matrix. This step could be parallelized on multi-processor systems for increased speed.) Once the SSR were calculated BP sequentially updated the $T-(m+1)h+1$ segments into one, two, and up to $(m-1)$ optimal breaks ($m$ partitions). The last step created a single optimal partition with $m$ breaks.

Let $SSR(\{T_{r,n}\})$ be the sum of squared residuals associated with the optimal partition containing $r$ breaks using the first $n$ observations. The optimal partition solves the recursive problem

$$SSR(\{T_{m,T}\}) = \min_{mh \leq j \leq T-h} \big[SSR(\{T_{m-1,j}\}) + SSR(j+1,T)\big]. \tag{1.4}$$

The full BP can handle multivariate linear regression with only some of the parameters undergoing breaks. This will not be considered as this thesis is confined to univariate time series.

It should not be surprising that Equation (1.4) bears some resemblance to Equation (1.8) below as they are both developments of the Fisher (1958) algorithm. Unless otherwise indicated we used the BP as implemented in the `R` package `strucchange` of Zeileis et al. (2002).

## 1.2 Application of CART to Time Series

In structural break detection and location methodologies the Classification And Regression Trees of Breiman et al. (1993) have been largely overlooked. Prior use of CART in time series analysis appeared in the paper of Cooper (1998), and the research report of Cappelli and Reale (2005). A variation of CART using smooth transition regressions was contained in a working paper which is now forthcoming as da Rosa et al. (2008). Apart from this small number of papers there appear to have been no other attempts to apply CART methodology as presented in Breiman et al. (1993), to the task of finding structural breaks in time series where the location and number of such breaks are

unknown. Rea et al. (2006), Cappelli et al. (2007), and Cappelli et al. (2008) contain work from this thesis. Thus the literature on the application of CART methodology to problems in time series analysis can only be described as sparse.

As the focus of this thesis is the application of regression trees to problems in time series analysis both Cooper (1998) and Cappelli and Reale (2005) will be reviewed in some detail.

### 1.2.1 Cooper 1998

The paper of Cooper (1998) applied CART methodology in a straight forward fashion by making what appears to be the first published application to time series analysis. She noted that with a regression tree there was no formal means of testing for the significance of the individual splits. (We note that the Smooth Transition Regression Tree procedure (STR-Tree) of da Rosa et al. (2008) may have solved this problem.)

Cooper stated that the use of an F-test was inappropriate because the regression tree maximized the improvement in fit from splitting the sample. The F statistic was also maximized by this procedure seriously distorting the critical values of the test. As there appeared to be no general solution to the problem of size distortion Cooper used Monte-Carlo simulation. Cooper did not consider the issue of the influence of serial correlation on the F-test.

Cooper's dataset consisted of monthly U.S. total industrial production plus 12 indices for individual industry groupings which made up the main index for the period from January 1923 to July 1991.

In our opinion Cooper did something unusual when growing her trees. In her Figures (3) and (4) she split some nodes on the time variable after other nodes had already been split using the lagged production variable. This appears to us to make the interpretation of the terminal nodes difficult. For example, in Cooper's Figure (3) node K was split into pre- and post-1945 but by this stage in the tree growing process three terminal nodes had already been created, nodes E, H, and J, some of which contained data from both periods. Similar types of comments can be applied to

the three terminal nodes in her Figure (4). The problem, as we see it, was that the terminal nodes contained observations which were not contiguous in time.

Cooper defended her choice of splitting on both the time variable and lagged growth. She claimed the non-linearity of the U.S. output included an interaction between time and lagged growth which splitting on time alone missed. She further claimed such things as short run fluctuations and persistence were found by her method.

### 1.2.2 Cappelli and Reale 2005

Cappelli and Reale (2005) introduced an application of the CART methodology of Breiman et al. (1993) as a structural break location method in time series. CART was a non-parametric procedure as it made no distributional assumptions about the data

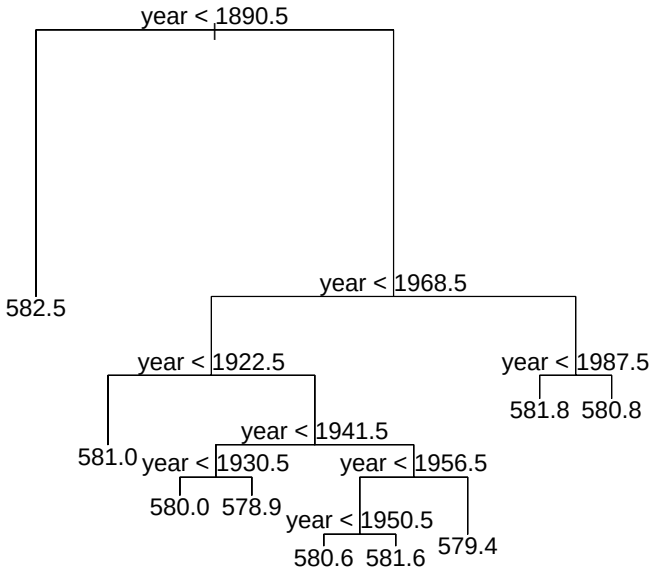


Figure 1.2: The regression tree for the Michigan-Huron data. The splitting variable is year AD and the response variable is mean lake height in meters above sea level.

or the residuals. Further, in the application of CART to univariate time series the time variable assumed the role of a predictor variable when, in fact, it was merely a counter. The lack of distributional assumptions and a true predictor variable led Cappelli and Reale to name this application Atheoretical Regression Trees (ART).

They applied this to an environmental time series of annual mean water level of the Lake Michigan and Huron system. The data set they used was an extension of the data set used by Fisher (1958). Fisher's data extended to 1955, while Cappelli and Reale's to 1999.

The advance of Cappelli and Reale's procedure over Fisher's method lay in finding the number of structural breaks in the data and the visual presentation of the findings. A regression tree for the Michigan-Huron lake level data is presented in Figure (1.2). As the lengths of the vertical lines connecting the parent with the child nodes indicate the reduction in the sum of squared residuals the relative importance of each of the breaks is immediately clear. If the researcher wished to prune this tree further it is clear which two nodes should be combined. By contrast Fisher's method required the number of breaks, $G$, to be specified in advance. In Fisher's method the data set was partitioned in to $G$ subgroupings with maximum homogeneity. The drawback was that the choice of $G$ lay with the researcher which required several trial partitions to be created. The best partition can now be selected with information criteria, such as the AIC, which were not available at the time to Fisher. It should be noted that because a regression tree only considers binary splits at each node and makes no reconsideration of the location of the splits at any point in the tree growing and pruning process the final partition of the data may be suboptimal.

With ART the number of structural breaks is determined by the data. ART builds a nested hierarchy which is then pruned back using the cost-complexity pruning originally proposed by Breiman et al. (1993) or an information criteria such as the AIC (Akaike, 1973), the BIC (Schwarz, 1978), the AICC (Hurvich and Tsai, 1989), or the RIC (Shi and Tsai, 2002). Because a regression tree fits piece-wise constant functions to the data, the penalty terms in the information criteria are only applied to the number of

terminal nodes. Because the model at a terminal node is that the data have a single mean the issue of further modelling of the data within a terminal node which would require further parameters does not arise.

The tree could also be pruned manually if required or by any other specified criteria. Once the final tree has been selected the terminal nodes represent the regimes which form the best partition of the data.

Cappelli and Reale compared the performance of the BP global optimal procedure against ART. While the BP procedure was known to be optimal it was computationally expensive. Their results showed ART reported very similar breakpoints to the those reported by the BP at a fraction of the computational cost.

Cappelli and Reale only applied ART to a single data set and no simulation study was carried out.

### 1.2.3 Atheoretical Regression Trees

In this section we give a brief overview of ART. When reading Breiman et al. (1993) it is clear that the authors did not expect their new methodology to be applied to time series. In the notation they developed for regression trees they used $t$, with or without subscripts, to denote the nodes of the tree and $T$ to denote the full tree. In time series usually $t$ denotes time and $T$ denotes the number of observations in the series. In what follows it should be clear from the context which usage is being employed.

If we are given a continuous response variable $Y$ and a set of $p$ predictors, $X_1, \ldots, X_p$, a regression tree will model the relationship between the response variable and the covariates by fitting piece-wise constant functions to the data. The points at which these piece-wise constant functions change are interpreted as candidate breakpoints in univariate time series. ART uses recursive partitioning which results in a partition of $Y$ based in the values of the $X_i;\ i = 1, \ldots, p$. In univariate time series there is only one exogenous predictor variable, time, which is strictly increasing. As mentioned above, the time variable is not a true predictor, rather it is a counter.

Breiman et al. (1993) examined both sum of squared residuals (SSR) and least

absolute deviation as measures of deviance in the tree. Least absolute deviation is less sensitive than SSR to extreme values and so may be a preferable measure when the data do have a number of extreme values.

The use of such a covariate in a least squares regression tree transforms the tree growing procedure into a recursive use of the Fisher (1958) method of contiguous grouping of $K$ elements into $G$ groups of maximum homogeneity where $G = 2$ at all non-terminal nodes.

Fisher identified two subclasses of the grouping problem. In the unrestricted case no conditions were imposed on the allowed partitions. This allowed the data to be sorted prior to analysis. In the restricted case additional conditions were imposed. Time series problems are of the restricted type as the data are ordered by time.

Fisher's method was computationally expensive. To group $K$ elements into $G$ mutually exclusive and exhaustive subsets of maximum homogeneity required $O(K^2G)$ steps. It only produced a single partitioning so it was advisable to vary $G$ to create several candidate partitions.

By contrast, at any node ART requires $O(n(t))$ steps to identify the best split where $n(t)$ is the number of elements in node $t$. Further, ART produces a hierarchical structure in the form of a binary tree. Final tree selection can be done automatically by a cost-complexity parameter as introduced by Breiman et al. (1993) or by an information criteria such as the Bayesian Information Criterion (BIC) (Schwarz, 1978). Alternatively, the researcher can prune the tree manually by using *a priori* knowledge.

ART is data-driven in the sense that the number of contiguous, maximally homogeneous groups and the locations of the breaks between the groups are uncovered by the procedure. Thus ART allows the data to show if any breaks exist, and if so, their location in time.

### 1.2.4 Growing a Regression Tree

In a short paper Hyafil and Rivest (1976) proved that the problem of finding the smallest decision tree consistent with a given data set was NP-complete. Thus it was

not feasible to search for an optimal tree except perhaps for very small data sets. The problem then became one of finding a tree which gave acceptable performance for reasonable computational cost.

Most tree construction methods use *greedy* algorithms. That is, the best split is taken at each step of the tree growing procedure and no reconsideration of splits, or back-tracking, is undertaken.

In this and the next section we describe the standard tree growing and pruning methods. However, the simulations (see Chapter 3) uncovered a weakness in the *greedy* algorithm used in growing a regression tree when applied to locating structural breaks in univariate time series. This will be addressed in Section (4.4).

In order to grow a regression tree we need a criterion which will give us a measure of node impurity. Two obvious candidates are the squared error and absolute deviation. For a node $t$ with $n(t)$ observations of the response variable $Y$ the squared error and absolute deviation are, respectively, given by

$$R(t) = \frac{1}{n(t)} \sum_{x_i \in t} (y_i - \bar{y}(t))^2 \tag{1.5}$$

and

$$R(t) = \frac{1}{n(t)} \sum_{x_i \in t} |y_i - \bar{y}(t)| \tag{1.6}$$

where

$$\bar{y}(t) = \frac{1}{n(t)} \sum_{x_i \in t} y_i \tag{1.7}$$

is the arithmetic mean of the elements of node $t$. The summation is carried out over $x_i \in t$ as the $x_i$ are the predictor variables which determine node membership, the $y_i$ are the response variables for which we are estimating node impurity. $R(t)$ is referred to as the risk function. While the use of least absolution deviation, Equation (1.6), is discussed in Breiman et al. (1993, pp255-264), it is the mean squared error, Equation (1.5), which is more commonly used.

The tree growing procedure seeks to split the node $t$ into two child nodes $t_L$ (left)

and $t_R$ (right) for which the sum

$$R(t_L) + R(t_R) = \frac{1}{n(t_L)} \sum_{x_i \in t_L} (y_i - \bar{y}(t_L))^2 + \frac{1}{n(t_R)} \sum_{x_i \in t_R} (y_i - \bar{y}(t_R))^2 \qquad (1.8)$$

is minimized. If we label this split $s$, the procedure seeks to maximize the quantity

$$\Delta R(s,t) = R(t) - R(t_L) - R(t_R). \qquad (1.9)$$

We could grow the tree until each node contains values of $y_i$ which are the same. In this case the node is said to be pure. The types of time series examined in this thesis rarely have two adjacent data points with the same value. Thus it is likely all terminal nodes would need to contain one element for them to be pure.

As Equation (1.5) is an estimate of the variance at node $t$, one way to regard the splitting criteria, Equation (1.9), is as a refinement of the probability model for the data. The whole tree growing process is then a hierarchical refinement of such models. The final regression tree yields a histogram as an approximation to the population function.

### 1.2.5 Pruning a Regression Tree

In the use of regression trees it is possible to grow a tree that fits the training data well but in fact has "over fitted" the tree. That is, the tree models well the specific features of the particular subset of $X$ but poorly reflects the population function. There is a need for a procedure akin to variable selection in regression.

The methodology introduced by Breiman et al. (1993, pp66–71 and pp279–296) is cost-complexity pruning which we now describe. Let $R_i$ be the risk function evaluated at each of the terminal nodes. Then the value of the risk function for the whole tree is

$$R(T) = \sum_i R_i(t).$$

We define the size of the tree to be the number of terminal nodes which we denote by $|\tilde{T}|$. Then we can define $R_\alpha$ as

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

where $\alpha$ is a real number called the complexity parameter. $\alpha$ is a penalty applied to the tree because of its complexity.

We search for a subtree $T(\alpha) \leq T_{\max}$ which minimizes the value of $R_\alpha(T)$. That is

$$R_\alpha(T(\alpha)) = \min_{T \leq T_{\max}} R_\alpha(T).$$

As there is a finite number of subtrees, by letting $\alpha$ increase, the pruning process produces a nested sequence of subtrees

$$T_{\max} \succ T_1 \succ \ldots \succ t_1$$

where $t_1$ denotes the root node and $\succ$ denotes the sequence of subtrees. The pruning process also generates a sequence of values for $\alpha$

$$0 = \alpha_1 < \alpha_2 < \ldots < \alpha_K. \tag{1.10}$$

Thus for $\alpha_k \leq \alpha < \alpha_{k+1}$, the tree $T_k$ is the smallest subtree which minimizes $R_\alpha(T)$.

Because a large tree gives apparent accuracy but may only be a consequence of the fact that the tree is optimized for the data used to grow it, what is required is a way to estimate the true error $R^*(T)$. This requires the use of test data drawn from the same population as the learning data. Breiman et al. (1993) suggest two ways.

1. Independent test data

2. Cross validation

In time series the data being analyzed are often the only realization observed, no other sample from the ensemble is available to act as independent test data.

The use of cross-validation with time series data involves some special problems, particularly in the search for structural breaks. We first outline the standard procedure for cross-validation, then discuss its problems.

In $V$-fold cross-validation the learning sample $L$ is divided into $V$ subsets $L_1, \ldots, L_V$ of as nearly equal size as possible. Breiman et al. (1993) recommend that the samples be kept balanced with respect to the response variable. To do this they suggest the

cases be put into levels based on the values of the response variable and that stratified random sampling be used to generate the $V$ balanced samples.

The $v$-th learning sample is denoted by $L^{(v)} = L - L_v$, the set $L_v$ is reserved for estimating the prediction error after the tree has been constructed.

The main tree is grown on $L$ and $V$ auxiliary trees are grown on the $L^{(v)}; v = 1, \ldots, V$ subsamples. We obtain $V + 1$ overly large trees; $T_{\max}$, $T_{\max}^{(1)}$, $\ldots$, $T_{\max}^{(V)}$. For each value of the complexity parameter, $\alpha$, let $T(\alpha)$ be the corresponding minimal cost-complexity subtree. Using the sequence of cost-complexity parameters from Equation (1.10) we now define

$$\alpha'_k = \sqrt{\alpha_k \alpha_{k+1}} \tag{1.11}$$

i.e. $\alpha'_k$ is the geometric mean of $\alpha_k$ and $\alpha_{k+1}$. We denote by $d_k^{(v)}(\mathbf{x})$, known as the decision function, the predictor corresponding to the tree $T^{(v)}(\alpha'_k)$. Then the cross-validation estimate for the prediction error is

$$\hat{R}^{cv}(T_k(\alpha'_k)) = \frac{1}{n} \sum_{i=1}^{V} \sum_{(\mathbf{x}_i, y_i) \in L_v} (y_i - d_k^{(v)}(\mathbf{x}_i))^2.$$

That is, we take the average value of the squared difference between the predicted response and the observed value.

Breiman et al. (1993, p234) also used a measure they called relative error which is given by

$$\hat{RE}^{cv}(T_k) = \frac{\hat{R}^{cv}(T_k)}{\hat{R}(\bar{y})}$$

where

$$\hat{R}(\bar{y}) = \frac{1}{N} \sum_n (y_n - \bar{y})^2$$

and

$$\bar{y} = \frac{1}{N} \sum_n y_n.$$

As we have $V$ trees to choose from, $\hat{R}^{cv}(T_k(\alpha'_k))$ is used to find the tree with the smallest subtree estimated error. Denote this by $T_{k_0}$. Finally, select the smallest tree $T_k^*$ such that

$$\hat{R}^{cv}(T_k^*) \leq \hat{R}^{cv}(T_{k_0}) + \hat{SE}(\hat{R}^{cv}(T_{k_0})) \tag{1.12}$$

where

$$\hat{SE}(\hat{R}^{cv}(T_{k_0})) = \sqrt{\frac{s^2}{n}}$$

and

$$s^2 = \frac{1}{n} \sum_{(\mathbf{x}_i, y_i)} \left[ (y_i - d_{k_0}^{(v)}(\mathbf{x}_i))^2 - \hat{R}^{cv}(T_{k_0}) \right]^2$$

and where $d_{k_0}^{(v)}(\mathbf{x}_i)$ is the predicted response according to the tree $T_{k_0}$. This method of selecting the final tree is sometimes called the one standard error rule (1SE).

When searching for structural breaks in time series, withholding part of the data as test data is inappropriate. One cannot be certain the withheld data has the same structure as the learning data set. With cross-validation a randomly selected subset of the data may have considerably different properties from the whole. For example, if a time series is best modeled as an AR(2) process, combining randomly selected data points into a new series will destroy the particular serial correlation in the original series.

Since the publication of Breiman et al. (1993) many papers have appeared on tree procedures. Esposito et al. (1997) reviewed six proposed tree pruning procedures in the context of classification trees. They noted that cost-complexity pruning with cross-validation made the assumption that the error rate of the $V$ trees $T^{(1)}(\sqrt{\alpha_i \alpha_{i+1}})$, $T^{(2)}(\sqrt{\alpha_i \alpha_{i+1}})$, $\ldots$, $T^{(V)}(\sqrt{\alpha_i \alpha_{i+1}})$ was the same as the true error rate of $T(\alpha_i)$. They pointed out this assumption was without theoretical foundation. Thus cross-validation may provide an estimate of the error rate with an unpredictable amount of bias.

In their experiments with 15 data sets Esposito et al. (1997) reported that the 1SE rule, Equation (1.12), was usually out-performed by simply using $\hat{R}^{cv}(T_{k_0})$. In fact, they concluded that in classification trees pruning did not generally increase the predictive accuracy of the pruned trees over $T_{\max}$.

Su et al. (2004) studied the use of information criteria to select an optimal tree from a sequence of pruned regression trees. They tested the CART algorithm against four information criteria; AIC, AICC, BIC, and RIC. They recommended using the BIC or RIC for routine tree selection. Their Figure (2) showed that the cost-complexity

pruning algorithm of CART gave an indistinct signal as to which tree was the "best". This can also be seen in Breiman et al. (1993, Figure 8.5 p235).

The BIC of Schwarz (1978) is, up to a constant, given by

$$\mathrm{BIC}(T_i) \propto n \cdot \log(\mathrm{SSE}(T_i)) + \log(n) \cdot |\tilde{T}_i|.$$

where

$$\begin{aligned} T_i &= \text{tree } i \text{ in the sequence } T_{\max} \succ T_1 \succ \ldots \succ t_1 \\ n &= \text{the number of observations} \\ \mathrm{SSE}(T_i) &= \text{ the sum of squared errors of tree } T_i \\ |\tilde{T}_i| &= \text{the number of terminal nodes in tree } T_i. \end{aligned}$$

This balances the conflicting requirements of a large tree with low SSE against a parsimonious tree with few terminal nodes. Minimizing the BIC selects the best subtree in the sequence which optimizes both requirements.

### 1.2.6 Stopping Rules

An alternative to growing a large tree and then pruning off nodes or subtrees which do not contribute to the tree's performance is to use a stopping rule. An example of a stopping rule is if we have a measure of node impurity $I$, which for regression trees is usually the variance of the node, set a threshold $\beta$ and do not split the node if

$$\max_s \Delta I(s,t) < \beta. \tag{1.13}$$

For multivariate trees stopping rules were generally unsatisfactory for two reasons.

1. If $\beta$ was set too low the tree was over-fitted and hence too large.

2. If $\beta$ was set too high it may be possible that there exists a node $t$ which fails the test of Equation (1.13) but if it had been split may yield nodes $t_L$ and $t_R$ one or both of which may yield large decreases in impurity on subsequent splits.

For univariate time series, as are being examined in this thesis, reason 2 is unlikely to occur in practice. This leaves us to deal with problem 1.

The obvious candidate for a stopping rule in univariate time series is the F-test. However, Cooper (1998) noted that the conventional F-distribution test for sample splitting was not appropriate for reasons outlined in Section (1.2.1).

We should note that da Rosa et al. (2008) use the Smooth Transition Regression presented by Granger and Teräsvirta (1993) in the context of regression trees. This allows the use of Lagrange Multiplier tests to be used as a test of hypothesis when deciding whether to split a node. See Hamilton (1994, p145) for a brief description of Lagrange Multiplier tests.

## 1.3 Structure of Thesis

The remainder of this thesis is set out as follows. Chapter (2) briefly looks at the theoretical performance of ART. In Chapter (3) the results of an extensive simulation study of ART are presented. Chapter (4) examines a weakness in the default tree growing and pruning procedures, proposes a method of solution for univariate time series we call Enhanced Temporal Pruning and evaluates its effectiveness. Chapter (5) is a straight forward application of ART to the Crest toothpaste data set.

The remainder of the thesis deals with the application of ART to a long standing problem in long memory time series, that of the duality between structural breaks and long memory. Chapter (6) reviews some of the relevant literature and presents several proposed models. Chapter (7) discusses a number of the estimators of the Hurst parameter $H$ and presents the results of an extensive simulation study into their properties. Chapter (8) outlines the empirical problem of the duality between breaks and long memory, proposes new methodology, presents a simulation study of the behaviour of ART when applied to simulated long memory series and looks at the new methods' behaviour in simulated data. Chapter (9) applies the new methodology to some financial data sets and compares it with some existing statistical tests for discriminating between true long memory and structural breaks. Chapter (10) does

the same for geophysical time series. Chapter (11) concludes.

## 1.4 Presentations and Publications

This section lists the publications and presentations which contain material from this thesis.

### 1.4.1 Refereed Publications

C. Cappelli, R. N. Penny, W. S. Rea and M. Reale; Detecting multiple mean breaks at unknown points with Atheoretical Regression Trees *Mathematics and Computers in Simulation* Vol. 78, No. 2-3, pp351-356, 2008

W. Rea, M. Reale, and J. Brown; Long Memory or Structural Breaks in Temperature and Proxy Time Series; *MODSIM 2007 International Congress on Modeling and Simulation*, pp. 3010-3016, L. Oxley and D. Kulasiri *eds.* Dec, 2007

G. Y. Zhang[1]; The Distributions of Change Points in Long Memory Processes *MODSIM 2007 International Congress on Modeling and Simulation*, pp. 3037-3043, L. Oxley and D. Kulasiri *eds.* Dec, 2007

Rea, W., Reale, M., Cappelli, C. and Brown, J.A. (2007). Modeling Long Memory Time Series: the Shihua Cave Speleotherms. In del Castillo, J., Espinal, A. and Puig P. (eds), *Proceedings of the 22nd International Workshop on Statistical Modeling*, IDESCAT Barcelona, pp 130-135

W. Rea, M. Reale, C. Cappelli and J. Brown; Identification of level shifts in stationary processes. *Proceedings of the 21st International Workshop on Statistical Modeling* pp 438–441. July 3-7 2006 Galway, Ireland

The poster paper on which this publication was based placed second in the poster paper competition.

[1]Both M. Reale and W. Rea were accidentally omitted as co-authors in the proceedings. They were listed as co-authors in the oral presentation.

### 1.4.2 Submitted Refereed Publications

W. Rea, M. Reale, L. Oxley and J. Brown; The Empirical Properties of Some Popular Estimators of Long Memory Processes; submitted to *Economic Letters*

W. Rea, M. Reale, C. Cappell iand J. Brown; Identification of Changes in Mean with Regression Trees: An Application to Market Research; submitted to *Econometric Reviews*

W. Rea, M. Reale, and J. Brown; Are long memory time series $H$-self-similar? An Empirical Study; submitted to *Statistical Methods and Applications*

W. Rea, M. Reale and J. Brown; Long Memory in Temperature Reconstructions; submitted to *Climatic Change*

### 1.4.3 Other Publications

C. Cappelli, W. S. Rea and M. Reale. The Application of Regression Trees to the Detecting of Multiple Structural Breaks in the Mean of a Time Series *Research Report UCMSD 2007/4* Department of Mathematics and Statistics, University of Canterbury

### 1.4.4 Conference Presentations

W. Rea, E. Mendes, L. Oxley, and M. Reale. Are stock market volatility series H-selfsimilar? *Workshop on Non-linear and Complex Systems Analysis.* Brisbane, Australia Sep 27-29, 2007.

W. Rea, E. Mendes, L. Oxley, and M. Reale. Realized Volatility: Long Memory or Shifting Means? An Empirical Study *New Zealand Econometric Study Group Meeting* 4-5 August 2007 Hamilton, New Zealand

W. Rea, M. Reale, and J. Brown; Do Long Memory Time Series Have Amnesia? *New Zealand Econometric Study Group Meeting* 5-6 August 2006 Dunedin, New Zealand

This presentation won a New Zealand Econometric Study Group – Reserve Bank of New Zealand award for econometric research.

### 1.4.5 Papers in Preparation

W. Rea, L. Oxley, M. Reale and E. Mendes; Long memory or shifting means? A new approach and application to realised volatility

# Chapter 2

# Theoretical Performance of ART

When using ART to discover structural breaks in data, if one were to use a stopping rule the obvious test to apply to determine the statistical significance of the candidate break is the F-test. However, Cooper (1998) pointed out that an F-statistic would be severely distorted and its significance relative to a conventional critical value cannot be used for two reasons.

First, as outlined in Section (1.2.4) the regression tree growing procedure locates candidate splits by maximizing the improvement in the sum of squared residuals. This will also maximize the F-statistic. Thus the calculated F value represents the most optimistic view of the data if one is looking for evidence of breaks and the most pessimistic if one believes the DGP is uniform.

Secondly, when calculating the F-statistic we wish to test against a null hypothesis that there are no breaks in the data. To do this in a general way the F-statistic is calculated conditional on having found a particular number of breaks in the data. With a regression tree the test is always a case of testing the null against an alternative of a single break at each node.

In the special case of the simulations in the following chapter, we know both the location and size of the breaks. This allows us to calculate a theoretical value of the F-statistic. This will give us some indication of how large a mean shift must be present in the data before we can expect ART to correctly locate it.

We consider the case of the noisy square wave simulations as presented in Section (3.3) below. This particular use of the F-test is taken from Zehna (1970, p530).

If $X_i; i = 1, 2, \ldots, 5$ are random variables associated with the five regimes then

$$\sum_{i=1}^{5} n_i \frac{(\bar{X}_i - \bar{\bar{X}})^2}{\sigma^2}$$

and

$$\sum_{i=1}^{5} \sum_{j=1}^{n_i} \frac{(X_{ij} - \bar{X}_i)^2}{\sigma^2}$$

where $\bar{X}_i$ is the mean of regime $i$ and $n_i$ is the number of observations in regime $i$, are independent random variables with $\chi^2_4$ and $\chi^2_{n-5}$ distributions respectively where $n$ is the total number of observations. If the null hypothesis is that there are no breaks, then the rejection rule at the $\alpha$-significance level is

$$\frac{\sum_{i=1}^{5} n_i \frac{(\bar{x}_i - \bar{\bar{x}})^2}{\sigma^2}}{\sum_{i=1}^{5} \sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_i)^2}{\sigma^2}} > \frac{4}{n-5} F_{n-5,4;1-\alpha}$$

or equivalently

$$\frac{n-5}{4} \frac{\sum_{i=1}^{5} n_i \frac{(\bar{x}_i - \bar{\bar{x}})^2}{\sigma^2}}{\sum_{i=1}^{5} \sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_i)^2}{\sigma^2}} > F_{n-5,4;1-\alpha}. \tag{2.1}$$

We will calculate one value in detail and then present a table of values for some of the parameter combinations used in the simulations.

We need to calculate

$$E\left[\frac{n-5}{4} \frac{\sum_{i=1}^{5} n_i \frac{(\bar{x}_i - \bar{\bar{x}})^2}{\sigma^2}}{\sum_{i=1}^{5} \sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_i)^2}{\sigma^2}}\right].$$

For the noisy square wave simulation presented in Section (3.3) with regime length 25

($n_i$) and break size of two standard deviations, $n = 125$ and

$$E[\bar{\bar{x}}] = 0$$

$$E[\bar{x}_1] = 0$$

$$E[\bar{x}_2] = 2$$

$$E[\bar{x}_3] = 0$$

$$E[\bar{x}_4] = -2$$

$$E[\bar{x}_5] = 0$$

$$E\left[\sum_{j=1}^{25}(x_{ij} - \bar{x}_i)^2\right] = 25; \text{ for } i = 1, 2, \ldots, 5.$$

$$E\left[\sum_{j=1}^{5} n_i(\bar{x}_i - \bar{\bar{x}})^2\right] = 200.$$

Thus the expected value of the F-statistic is

$$E[F_{120,4,0.05}] = \frac{120}{4}\frac{200}{125} = 48$$

Table (2.1) presents some values of the F-statistic for a few parameter combinations from the noisy square wave simulations. As the degrees of freedom in the denominator were always greater than or equal to 120 we could use $F_{\infty,4;0.05}$. For comparison purposes $F_{120,4;0.05} = 5.66$ and $F_{\infty,4;0.05} = 5.63$.

We can calculate the theoretical value of the break size in standard deviations which will result in an F-value of 5.66. From Equation (2.1) we obtain

$$\frac{\sum_{i=1}^{5} n_i \frac{(\bar{x}_i - \bar{\bar{x}})^2}{\sigma^2}}{\sum_{i=1}^{5}\sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_i)^2}{\sigma^2}} = \frac{4}{n-5} F_{n-5,4;0.05}.$$

Table 2.1: The calculated F-statistic values are for four break simulations.

| Regime Length | 25 | 100 | 400 |
|---|---|---|---|
| Breaksize | | | |
| 2.00 | 48.0 | 198.4 | 448.4 |
| 1.00 | 12.1 | 49.6 | 112.1 |
| 0.50 | 3.0 | 12.4 | 28.0 |
| 0.25 | 0.85 | 3.48 | 7.85 |

Using the following values

$$\bar{\bar{x}} = 0$$
$$\bar{x}_1 = 0 = \bar{x}_3 = \bar{x}_5$$
$$\bar{x}_2 = 2 = -\bar{x}_4$$
$$E[(x_{ij} - \bar{x}_i)^2] = 1; \text{ for } i = 1, 2, \ldots, 5.$$

This becomes

$$\frac{2n_i(\bar{x}_2)^2}{5n_i} = \frac{4}{n-5}F_{n-5,4;0.05}$$
$$\bar{x}_2^2 = \frac{10}{n-5}5.66$$
$$\bar{x}_2 = \sqrt{\frac{56.6}{n-5}}.$$

The results are listed in Table (2.2). These are the theoretical break sizes below which we would not reject the null hypothesis of no breaks at the five percent level.

Thus for the noisy square waves and regime length 25 data points the smallest mean shift we can expect ART to reliably find would be about two-thirds of a standard deviation. Whereas for a regime length of 400 data points the smallest mean shift would be about one-sixth of a standard deviation.

Table 2.2: The theoretical values of the break size below which we would not reject the null hypothesis of no breaks in the data using an $F$-test.

| Regime Length | Breaksize |
|---|---|
| 25 | 0.687 |
| 36 | 0.569 |
| 49 | 0.486 |
| 64 | 0.424 |
| 81 | 0.376 |
| 100 | 0.338 |
| 121 | 0.307 |
| 144 | 0.281 |
| 169 | 0.260 |
| 196 | 0.241 |
| 225 | 0.225 |
| 256 | 0.211 |
| 289 | 0.198 |
| 324 | 0.187 |
| 361 | 0.177 |
| 400 | 0.168 |

# Chapter 3

# The ART of the Mean Simulations

## Introduction

ART was introduced as an application of the CART methodology of Breiman et al. (1993) by Cappelli et al. (2008) to the problem of locating mean shifts in time series where the number and location of the shifts were unknown. Cappelli et al. (2008) built on the work of Cappelli and Reale (2005) and Cappelli et al. (2007). The suitability of ART as a method to detect structural breaks in time series was assessed by simulation.

There were several specific questions to be addressed before ART could be used in applied research into problems in time series analysis. These are:

1. As ART fits piecewise constant functions to data, does ART discover breaks in, or impose breaks on, time series?

2. What is the effect of serial correlations on ART's performance in detecting structural breaks?

3. What is the effect of non-Gaussian noise structures on ART's performance in detecting structural breaks?

4. Given that observations in time series are, in general, non-interchangeable, can cross-validation be used in tree selection?

5. Given that ART is usually implemented using a least squares impurity function, is ART robust to outliers in the data?

6. Is it possible to obtain a confidence interval for the breaks?

In this chapter we explain simulations used to assess ART, present the results and comment on ART's strengths and weaknesses. Individual simulations involved up to 1,560 different combinations of parameters and unless otherwise stated, 1,000 replications of each parameter combination were run. To present the summarized data in tables typically required 15 to 20 pages per simulation. Thus most simulations are not presented as numerical tables of results for reasons of space. Instead, the results are presented as graphs. Because many of the graphs are surfaces in three dimensions we have often rotated them from their default orientations to give a better view of the surface. The reader should take careful note of the two horizontal axis to ensure correct interpretation. We draw attention to this where appropriate.

We use the term "spurious break" to describe the event when a procedure reported a structural break at a different location than the one put in the simulated series. There are two sources of spurious breaks. The first arises from the fact that a purely random process can generate data which appears to possess a break within the criteria of the test. The second arises because some tests, particularly the BP, report a confidence interval for the location of the break. If a true break lies outside the reported confidence interval, the reported break appears spurious.

All simulations were run using the statistical software package `R` produced by the R Development Core Team (2005). The series were generated using standard functions such as `arima.sim`, `rgeom`, `rnorm`, and `ts`. ART was implemented using functions from the contributed package `tree` of Ripley (2005).

The remainder of the chapter is set out as follows. Section (3.1) covers simulations for series with a single break. Section (3.2) covers the simulations for series

with two equal and opposite sized breaks. Section (3.3) covers simulations for series with four breaks in the form of a square wave. Section (3.4) covers simulations for series with heteroskedasticity. Section (3.5) compares the effectiveness of leave-one-out cross-validation with cost-complexity and the BIC as pruning techniques in series with varying levels of serial correlation. Section (3.6) examines the robustness of ART to outliers in the data. Section (3.7) examines the use of the bootstrap to establish confidence intervals for the reported breakpoints. Section (3.8) reports comparisons of compute times among the methods. Section (3.9) presents the conclusions. Chapter (4) addresses a problem observed in the simulations where ART reported an incorrect split in the root node. The simulations are not presented in chronological order. Some choices of parameters assume results which are presented later.

## 3.1 Series with a Single Breakpoint

In this section we explain the simulations and present the results for simulations which only have one break in the series. These simulations allow us to assess ART's ability to find a break in the absence of any interaction between the breaks or any other special effects which may influence the detection of the break such as were discovered in the noisy square wave simulations presented in Section (3.3) below.

### 3.1.1 Break at the Mid-point of Uncorrelated Data

A set of simulations were run with series of uncorrelated observations drawn from an N(0,1) population with a single breakpoint at the midpoint of the series. The series therefore had two regimes. In most simulations there were 16 regime sizes, $5^2$ to $20^2$ observations in length. The axis labeled "Regime Number" is non-linear in scale. Regime number 5 corresponds to series in which the regime length was $5^2$ (or 25) observations and similar up to regime number 20 which corresponds to series in which the regime length was $20^2$ or 400 observations. The break sizes ranged from 0.05 to 2 standard deviations in steps of 0.05 standard deviations. Default cost-complexity

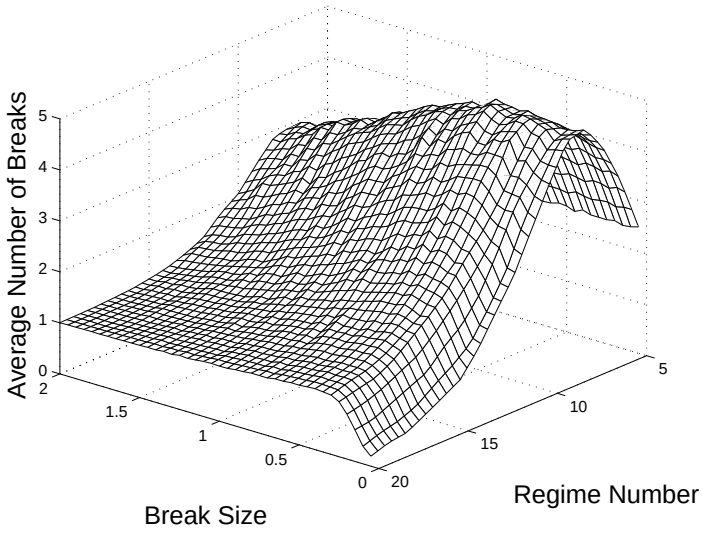


Figure 3.1: Average total number of candidate breaks reported by ART in simulated series with a single break at the mid-point and using default cost-complexity pruning.

pruning was used.

The results for the average total number of reported breaks are presented in Figure (3.1).

As could be expected of a statistical procedure based on OLS, ART's power increased with the length of the series. The results showed ART performed well when the regime length was long and the break size was large.

This and subsequent sets of simulations exposed a problem. ART reported substantial numbers of spurious breaks when the regime length was small and the default cost-complexity pruning was used. Su et al. (2004) reported that cost-complexity pruning as developed in standard CART methodology was inferior to several other pruning criteria. They found the BIC and RIC to give the best results. This is addressed further in Section (3.3.3). For long regimes few spurious candidate breaks were reported for small break sizes and no spurious breaks for large break sizes. Often for these series the tree did not need any further pruning.

For comparison purposes the BP was run on some series hence the horizontal axis

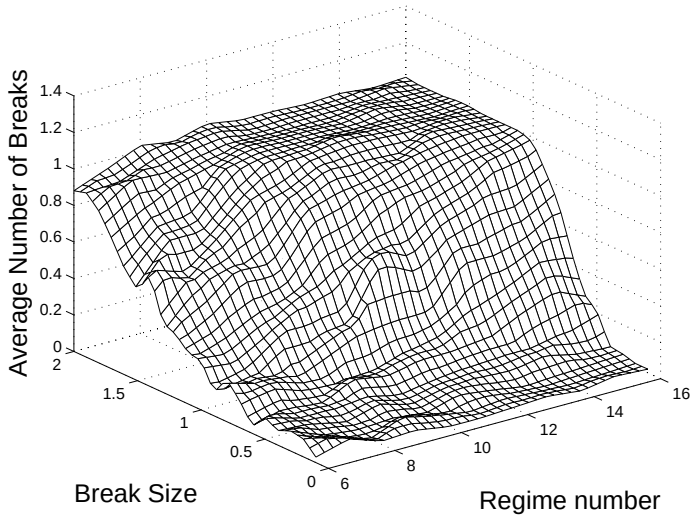


Figure 3.2: Average total number of breaks reported by the BP in simulated series with a single break at the mid-point.

"Regime Number" is shorter than the corresponding axis for ART. The results are presented in Figure (3.2). Note the orientation of the "Regime Number" axis compared to Figure (3.1). The BP rarely overfit the series, but often underfit in the short regimes and small break sizes.

### 3.1.2 CUSUM Tests

The CUSUM test is useful to compare with other tests which seek to locate the break. In this section we report the results of the CUSUM tests on the same set of simulations outlined in the previous section. Details of the CUSUM test were outlined in Section (1.1.1). To very briefly recap, the null hypothesis, $H_0$, of a CUSUM test is that there are no structural breaks in the series. The alternative, $H_\alpha$, is that there are one or more structural breaks present. The test reported, among its results, a $p$-value which could be used to accept or reject the null hypothesis. The results are presented in Figure (3.3). Note the orientation of the axes compared to Figures (3.1) and (3.2).

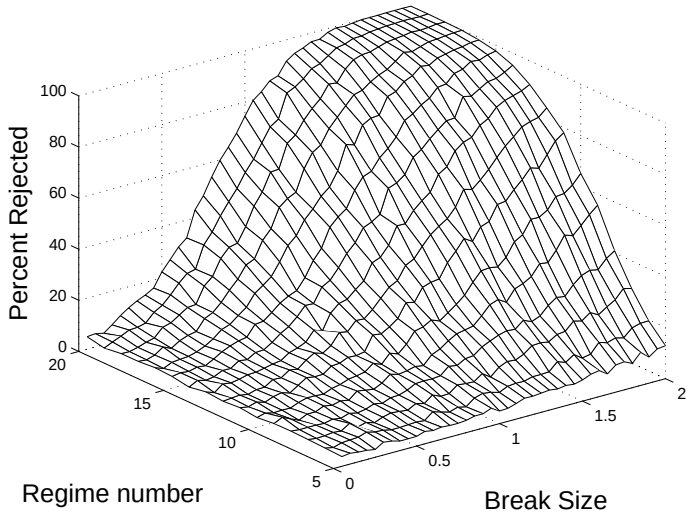


Figure 3.3: The percentage of series with a single break at the mid-point for which the null hypothesis was rejected at the five-percent level by a CUSUM test.

As the regime length increased the CUSUM test rejected the null hypothesis more frequently than in short series with the same break sizes. Similarly, for larger break sizes the CUSUM test rejected the null hypothesis more frequently than series with smaller break sizes. The minimum break size tested was 0.05 standard deviations. The minimum rejection level appeared to be about five percent, consistent with the chosen level of significance.

Comparing Figure (3.3) with Figure (3.1) and noting the orientation of the horizontal axes in these graphs are different, ART consistently reported breaks in parameter regions where the CUSUM test had a low rejection rate of the null hypothesis. Note particularly CUSUM's ability to reject $H_0$ declined quickly when both the regime length and the break size decreased. ART reported a break when the break size was large even in short series. Thus ART appeared to be over-reporting breaks compared to the CUSUM test.

### 3.1.3 Location of Break

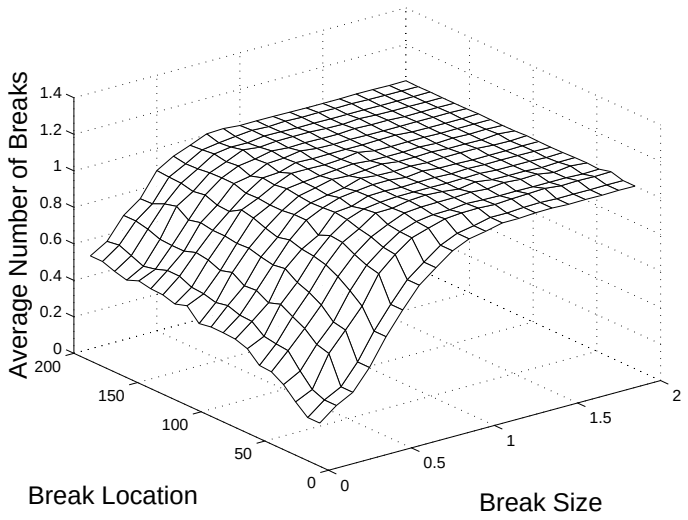


Figure 3.4: Average number of breaks reported by ART for a series of 400 data points in length as a function of break size and location with BIC tree selection.

In this subsection we examine the effect of the location of the break on ART's ability to correctly report a break. We compare these with one set of results from the BP. The series lengths were fixed at 100, 400, and 1600 data points and the breakpoint was moved within the first half of the series. This gave two regimes of unequal length. The step sizes for moving the location of the break were 5, 10, and 20 data points for the 100, 400, and 1600 data point series respectively. We assumed the ability of ART (and the BP) to detect a break was symmetrical so only moved the location of the break within the first half of the series.

The results for the 400 data point series are presented in Figure (3.4). The remainder are omitted for reasons of space.

A problem arose in these simulations with long series when the break was very close to the start of the series. In the growing of the maximal tree the software exceeded the maximum depth allowed. The `R` package `tree` can only grow trees to a maximum depth

of 31 layers. While this seems plenty, when the break was at the start of the series it grew the tree essentially with one leader branch splitting to the right and leaving small terminal nodes to the left. Thus 31 layers were exceeded before reaching the end of the series. We had to adjust the `mincut` parameter, which is explained in Section (3.1.7), in the 400 and 1600 data point series to run the simulations. This means these simulations are not directly comparable with most others in this chapter for which a value of `minicut`=5 was used.

These simulations showed that the ability of ART to find real breaks in the series was not dependent on the break location in the series of lengths 100 and 400. The break could be up to the value of the `mincut` parameter within the end of the series and ART was still able to locate it. The determining factor in finding the break was its size.

In the series of length 1600, ART showed reduced effectiveness in locating breaks near the start of the series. This could be seen with the break location at data point 20 and to a lesser extent at data point 40. Otherwise its ability to correctly locate the break depended on the break size as in the other two cases.

### 3.1.4 Series with AR(1) Correlations

The purpose of the AR(1) simulations presented in this section and the AR(2) and MA(1) simulations presented in Sections (3.1.5) and (3.1.6) respectively was to assess ART's ability to correctly locate breaks in the presence of serial correlations.

There is a convention used in which causal AR models are often referred to as stationary. If an AR(p) model is non-stationary when regressed on pass values of a time series it can always be converted to a stationary model by instead regressing on future values. In practice time series analysts do not build AR models regressed on future values and the distinction between causal models, which are only regressed on past values, and stationary models is often neglected, generally with no serious consequences. Chan (2002, pp26-28) gives a short summary of the differences between causal and stationary models while Brockwell and Davis (2006, pp77-89) gives a more

elaborate description. Because it is common practice to use the term stationary when causal would be more correct, in what follows we will refer to causal AR models as stationary but the reader should be aware this terminology is somewhat loose.

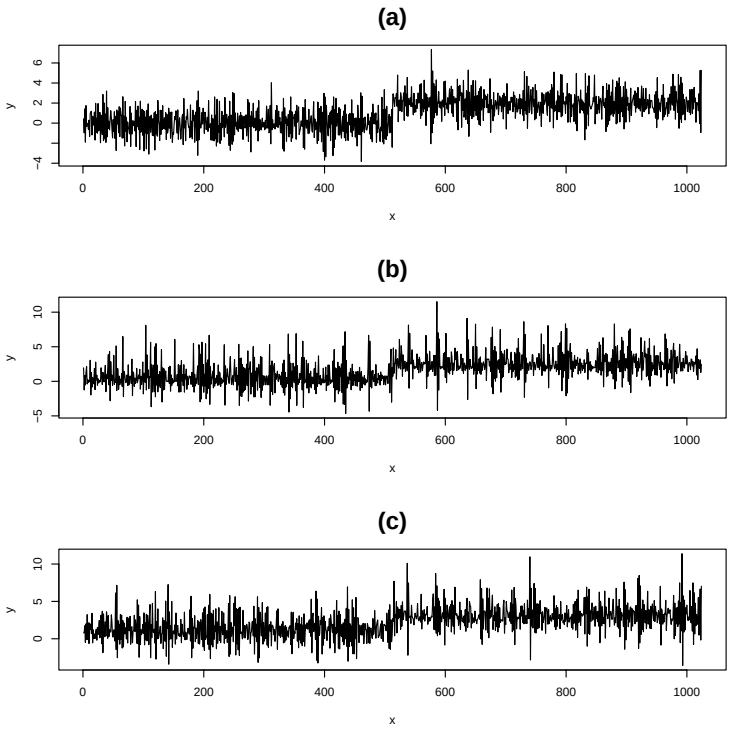


Figure 3.5: Three AR(1) series with $\phi = -0.65$, break size 2 and (a) Gaussian, (b) geometric and (c) gamma noise structures.

An autoregressive process of order 1 (AR(1)), $X_t$, can be expressed as

$$X_t = \phi X_{t-1} + Z_t \tag{3.1}$$

where $\phi$ is the AR(1) parameter and $Z_t$ is a random noise process.

Provided $|\phi| < 1$ the AR(1) process is stationary and invertible and we can write it as an infinite order MA process

$$X_t = \sum_{i=0}^{\infty} \phi^i Z_{t-i}.$$

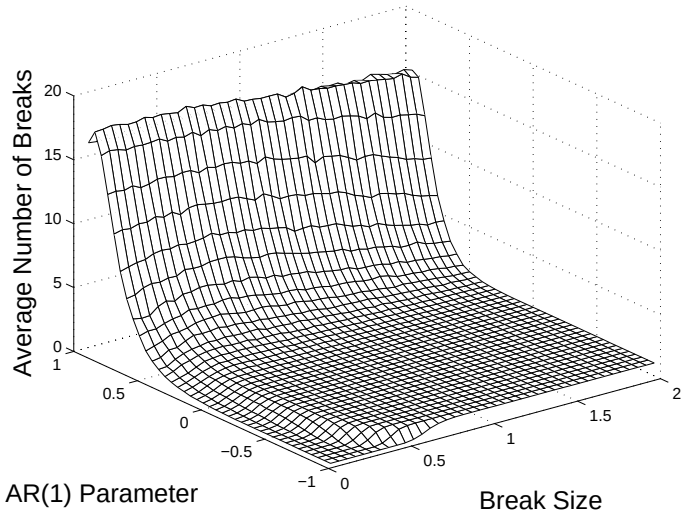


Figure 3.6: Average total number of breaks reported by ART as a function of AR(1) parameter and break size with Gaussian noise and BIC tree selection.

From this it is a simple matter to show that the variance of the process is

$$\begin{aligned} \sigma_X^2 &= \sigma_Z^2 \sum_{i=0}^{\infty} \phi^{2i} \\ &= \frac{\sigma_Z^2}{1-\phi^2} \end{aligned} \tag{3.2}$$

with the standard deviation obtained in the usual way. Thus for values of $|\phi|$ close to 1, the variance of the series will be much larger than the variance of the noise terms. This should be borne in mind when interpreting the performance of ART (or the BP) on series with serial correlations.

While it would have been a simple programming task to adjust the break size to correspond to the variance of the process rather than the variance of the noise terms, it was not clear that this would better reflect real time series. Instead, we chose to express all break sizes in units of the input noise terms.

All series were 1024 observations in length with a single break at the mid-point. The break was added after the full series was generated. We used three noise structures;

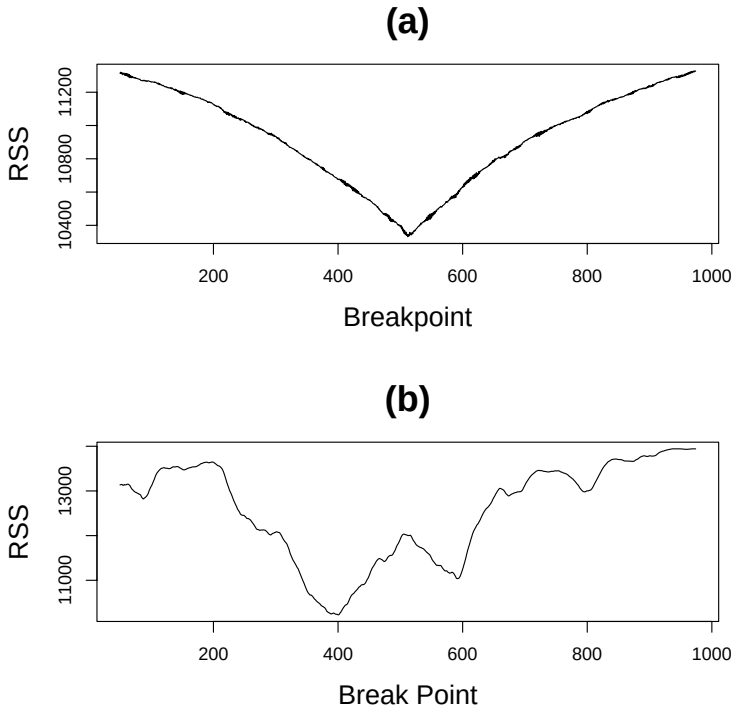


Figure 3.7: Residual sum of squares as a function of candidate breakpoint location for a series of length 1024 data points, break size two standard deviations, Gaussian noise and (a) $\phi = -0.95$ and (b) $\phi = 0.95$.

Gaussian, gamma and geometric. A plot of one realization of each of these three noise structures is presented in Figure (3.5).

**Gaussian Noise**

The Gaussian noise was drawn from an N(0,1) distribution. The results are presented in Figure (3.6).

When comparing Figure (3.6) with Figure (3.1) we should note the uncorrelated case is when the AR(1) parameter $\phi = 0$ and that the regime lengths are 512 data points in Figure (3.6) and a maximum of 400 data points in Figure (3.1). For $-0.5 < \phi < 0$ ART's ability to locate a break was enhanced. When $\phi < -0.5$ ART's performance

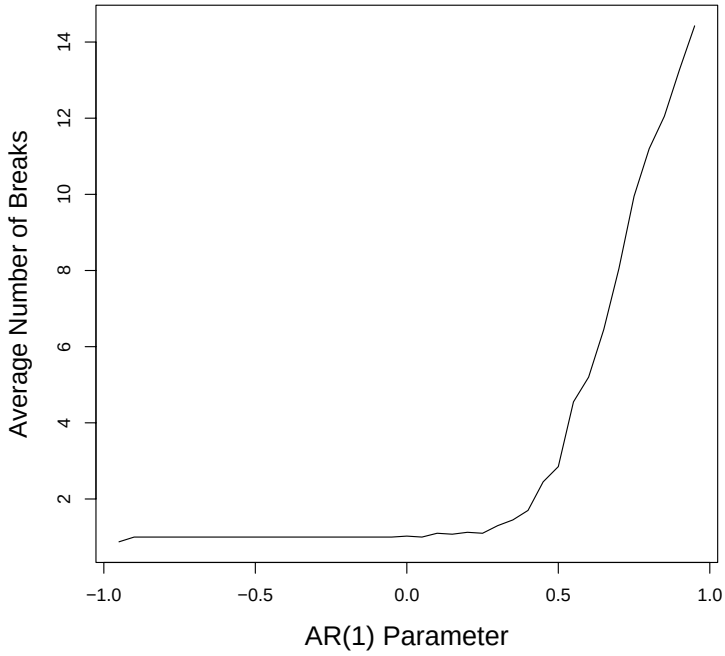


Figure 3.8: Average total number of breaks reported by the BP as a function of AR(1) parameter with Gaussian noise and a break size of two standard deviations.

was worse at small break sizes. In the case $\phi < -0.5$ the variance of the series was much larger than the variance of the noise terms. It can be shown (see Chatfield, 2004, p115) that the spectral density of an AR(1) process is given by

$$f(\omega) = \frac{1}{\pi(1 - 2\phi\cos\omega + \phi^2)}. \tag{3.3}$$

Thus for $\phi < 0$ the power is concentrated in the high frequency regions. This means the series oscillates rapidly about the mean. This generates a clear minimum RSS as can be seen in panel (a) of Figure (3.7).

Figure (3.7) presents the RSS as a function of candidate break location as seen by ART when deciding where to split the root node. With $\phi = -0.95$ the minimum was clearly at the location of the true break. It should be intuitively clear that after splitting the root node no further splits would be made in this series. This corresponds

to the large flat parameter region in the front of Figure (3.6). Part of the reduced effectiveness of ART in reporting breaks when $-0.95 < \phi < -0.5$ could be attributed to the fact that the effect of a large value of $Z_t$ decayed only slowly with time.

With $\phi = +0.95$ (see panel (b) of Figure 3.7) the true breakpoint was not close to a local minimum. In this particular case the true break would probably not be reported. This corresponds to the parameter region at the back of Figure (3.6). It should be intuitively clear that after splitting the root node (here very close to data point 400) there would be a number of other places where ART would split the series and report candidate breaks in attempting to minimize the RSS.

Figure (3.8) shows the BP reported similar results to ART. Here the break size was two standard deviations in terms of the noise series. Once the AR(1) parameter, $\phi$, exceeded about 0.25 the procedure began to report spurious breaks in rapidly increasing numbers. The comparable results in Figure (3.6) are the parameter region with break size two on the far right hand side of the graph.

The results for the gamma and geometrically distributed noise were nearly indistinguishable from the Gaussian noise and are not presented here for reasons of space.

**Gamma Noise**

In panel (c) of Figure (3.5) we have plotted a realization of an AR(1) process with gamma noise.

The choice of gamma parameters was $\alpha = 2$ and $\beta = 1$. The gamma distribution is given by

$$f(x; \alpha, \beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} \tag{3.4}$$

with mean $\mu = \alpha\beta = 2$ and variance $\sigma^2 = \alpha\beta^2 = 2$.

We examined the break locations from the simulations for break size 0.4 and AR(1) parameter -0.3. This parameter combination was in the equivalent region for the gamma distribution simulations to the large flat area in the foreground of Figure (3.6). These results are presented as a histogram in Figure (3.9). ART correctly determined that the series had a single break. However, the distribution of breakpoints covered a wide

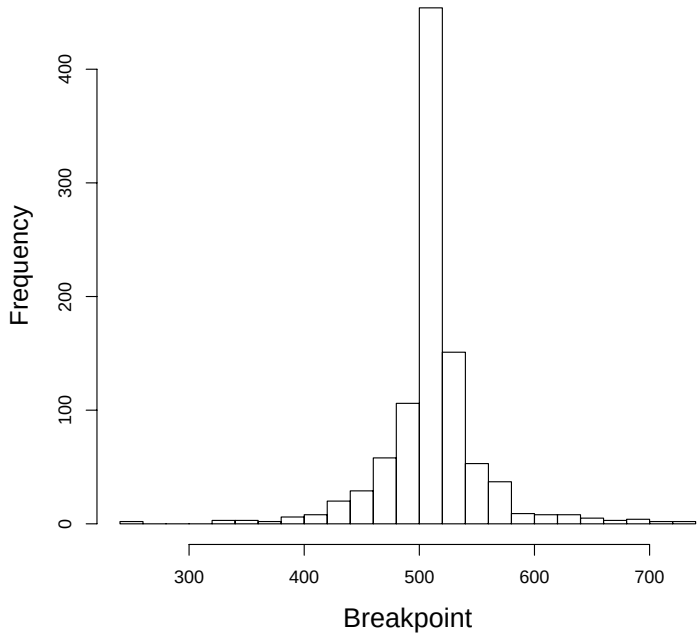


Figure 3.9: The distribution of breaks reported by ART for AR(1) parameter = -0.3, break size 0.4 with gamma distributed noise where $\alpha = 2, \beta = 1$ and BIC tree selection.

range from less than data point 300 to higher than data point 700.

### Geometric Noise

The choice of parameter for the geometric distribution was $\theta = 0.5$[1]. For a geometric distribution

$$g(x;\theta) = \theta(1-\theta)^{x-1} \text{ for } x = 1, 2, 3, \ldots \tag{3.5}$$

$$\mu = \frac{1}{\theta}$$

[1] $\theta$ is the standard notation for the parameter of the geometric distributions. It is also the standard notation for the MA(q) parameters with subscripts if $q > 1$, and in the literature on long memory $\theta$ is often used to denote a vector of parameters to be estimated. It should be clear from the context which usage is being employed.

and

$$\sigma^2 = \frac{1-\theta}{\theta}.$$

Thus for $\theta = 0.5$, $\mu = 2$, and $\sigma^2 = 1$. The results were almost indistinguishable from the gamma noise results and are omitted for reasons of space.

### 3.1.5 Series with AR(2) Correlations

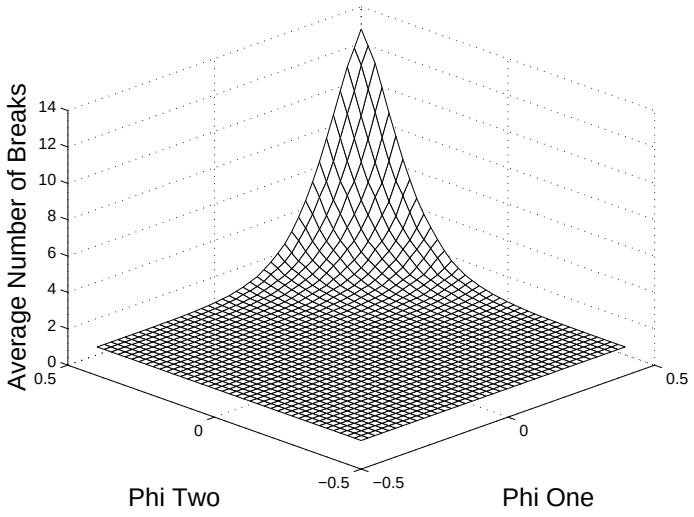


Figure 3.10: Average total number of breaks reported by ART as a function of the two AR(2) parameters with break size two, Gaussian noise and BIC tree selection.

An autoregressive process order 2 (AR(2)), $X_t$, can be expressed as

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t \tag{3.6}$$

where $\phi_1, \phi_2$ are constants and $Z_t$ is a noise term as before.

In these simulations the series lengths were 1024 data points, all break sizes were two standard deviations in terms the input noise series and a single break at the mid-point of the series. Values of $|\phi_1|, |\phi_2| > 0.5$ were not considered. The variance of an AR(2) process depends on the values of $\phi_1$ and $\phi_2$ but this does not appear to have been an issue in this set of simulations.

The results are presented in Figure (3.10).

It can be shown (see Diggle, 1990, p77) that the spectrum of an AR(2) process is given by

$$f(\omega) = \frac{1}{\left(1 - \sum_{l=1}^{2} \phi_l \cos l\omega\right)^2 + \left(\sum_{l=1}^{2} \phi_l \sin l\omega\right)^2}. \tag{3.7}$$

In the range of values for $\phi_1$ and $\phi_2$ which we considered, the power was concentrated in the lower frequencies when both $\phi_1$ and $\phi_2$ were close to 0.5. As with the AR(1) simulations, the presence of low frequencies in the data was interpreted by ART as structural breaks. This resulted in a decreased ability locate the actual break in a series. This can be seen in the parameter region at the back of Figure (3.10) in which many spurious breaks were reported.

### 3.1.6 Series with MA(1) Correlations

A moving average process order 1 (MA(1)), $X_t$, can be written as

$$X_t = Z_t + \theta Z_{t-1} \tag{3.8}$$

where $\theta$ is the MA(1) parameter and $Z_t$ are noise terms. MA(1) processes are always stationary and are invertible if $|\theta| < 1$. We did not consider non-invertible models. The variance of an MA(1) process is given by

$$\sigma_x^2 = (1 + \theta^2)\sigma_z^2. \tag{3.9}$$

All series were 1024 observations in length with a single break at the mid-point. The break was added after the full series was generated. A single realization of an MA(1) series with three different noise structures are presented in Figure (3.11).

The results for series with geometric noise are presented in Figure (3.12). The results for the gamma series were almost indistinguishable from series with geometric noise while for the series with Gaussian noise slightly lower numbers were reported when $\theta > 0$. Thus for reasons of space the Gaussian and geometric series are omitted.

It can be shown (see Diggle, 1990, p74) that the spectrum of an MA(1) process is given by

$$f(\omega) = 1 + 2\theta \cos\omega + \theta^2. \tag{3.10}$$

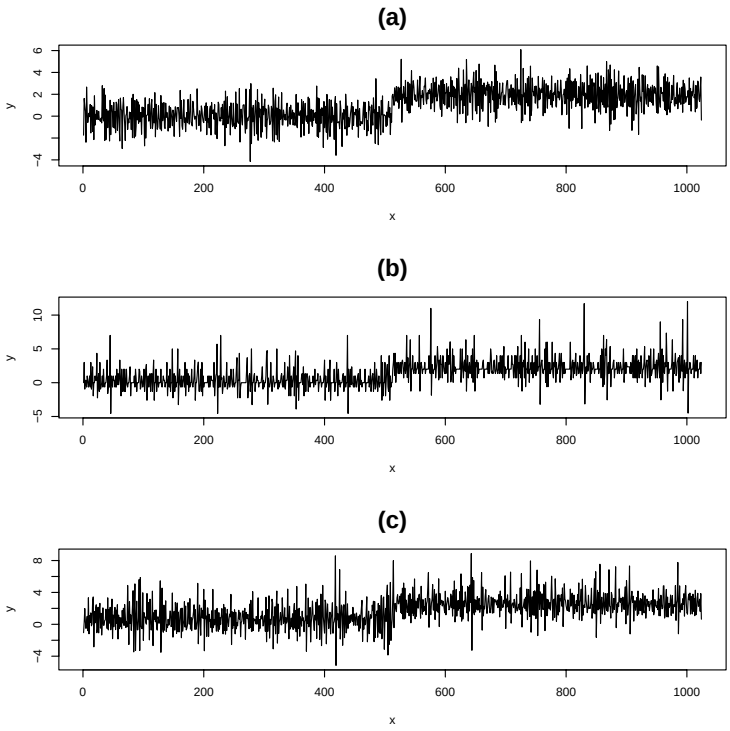


Figure 3.11: Three MA(1) series with a single break at the mid-point, $\theta = -0.65$ and (a) Gaussian, (b) geometric, and (c) gamma noise terms.

This has some features in common with the AR(1) spectrum (Equation 3.3). When $\theta < 0$ the power is concentrated in the higher frequencies. For $\theta > 0$ the power is concentrated in the lower frequencies. It differs in that the MA(1) spectrum is bounded as $\theta \to -1$ or $\theta \to 1$. The AR(1) spectrum tends to infinity as $\omega \to 0$ and $\phi \to 1$.

As in the AR(1) case ART tended to report spurious breaks when low frequencies were present in the data. Low frequencies were present in the parameter region toward the back of Figure (3.12). The rise in total breaks reported was much less dramatic than in the AR(1) case as the spectrum was bounded for MA(1) series.

For large break sizes, values of $\theta$ close to one only slightly impaired ART's ability to correctly locate the break. The problem became progressively more serious as the break size decreased.

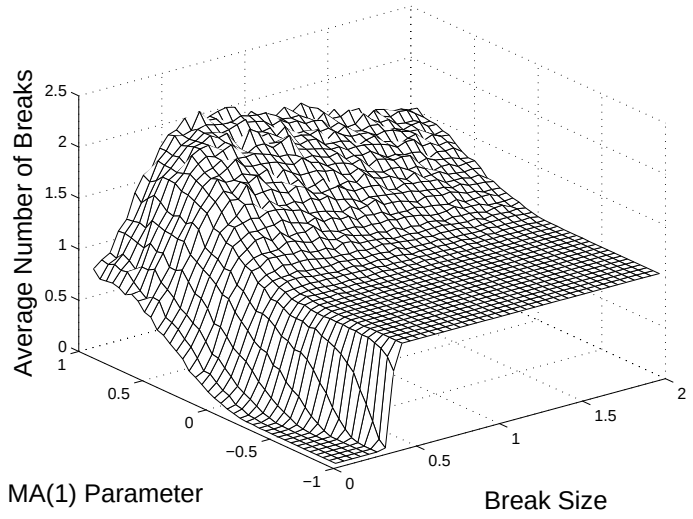


Figure 3.12: Average total number of breaks reported by ART as a function of MA(1) parameter and break size with geometrically distributed noise where the geometric distribution parameter was $\theta = 0.5$ and BIC tree selection.

### 3.1.7 Minimum Cut Size

da Rosa et al. (2008) observed that for small samples the tree growing and pruning algorithms produced trees with much higher numbers of terminal nodes than expected. They referred to this as over-fitting. In this section we look at the effect of the `mincut` parameter on the final tree. The `mincut` parameter was used in deciding whether to split a node. If the candidate split would produce a node with less than `mincut` observations then the node was not split.

As the tendency to select too large a tree was most evident in the simulations with regime length 36 (i.e. regime number 6), we selected that size for testing. The simulations were run as in the single break simulations in Section (3.1.1) and `mincut` values were set to five, 10, and 15 data points. A value of five was the default for the software we used.

The results are presented in Figure (3.13). These showed that raising the value of

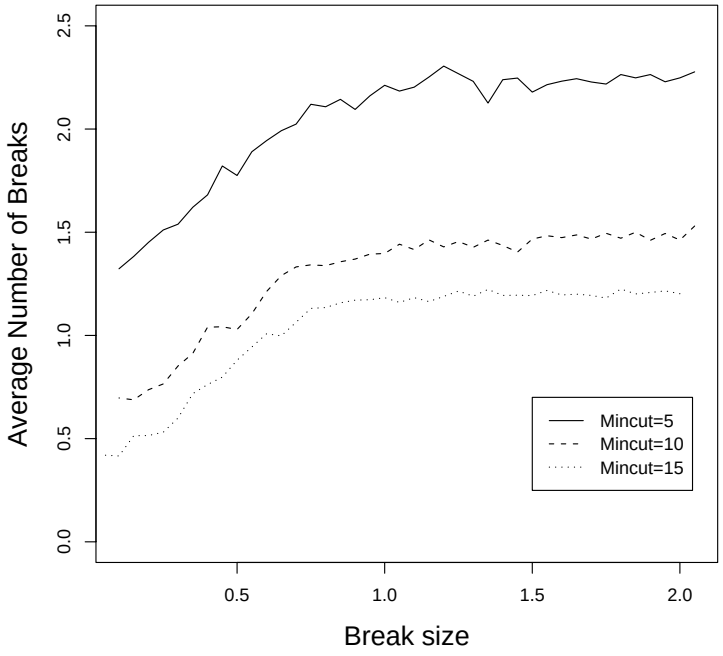


Figure 3.13: Average total number of breaks reported by ART for a series of length 72 and one break at the mid-point for various values of the `mincut` parameter and BIC tree selection.

`mincut` reduced the number of reported breaks. Raising `mincut` from five to 10 resulted in an average reduction of approximately 0.75 breaks per series. Raising it further to 15 resulted in an additional reduction of approximately 0.25 breaks per series. Given the tendency for ART to overfit breakpoints in short series this was a worthwhile reduction. There was also a small reduction in the amount of CPU time required for tree construction as shown in Table (3.1).

Given that ART only had a problem with short series, changing the `mincut` parameter and re-running the analysis may be beneficial if a lot of terminal nodes remain on the final tree after BIC pruning.

Table 3.1: CPU time required to generate and analyze 40,000 series of length 72 for various values of `mincut`.

| Mincut | CPU Seconds |
|---|---|
| 5 | 2893 |
| 10 | 2701 |
| 15 | 2540 |

## 3.2 Two Equal and Opposite Sized Breaks

An anonymous referee for Econometric Reviews suggested we examine ART's ability to correctly locate breaks when there were two offsetting breaks of equal size. The referee described this setup as "notoriously difficult". The results of these simulations are presented in Figures (3.14) and (3.15).

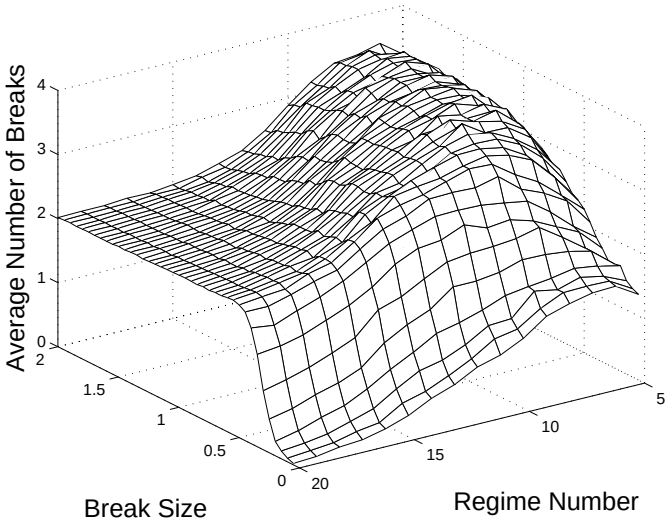


Figure 3.14: Average total number of breaks reported by ART in the two offsetting break simulations with Gaussian noise and BIC tree selection.

Figure (3.14) presents the average total number of breaks reported by ART. This

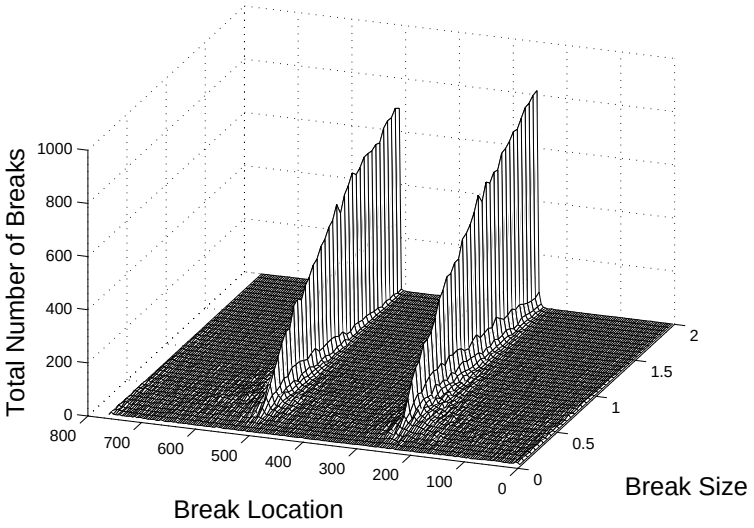


Figure 3.15: The locations of the breaks found for the series with regime length 256 data points and Gaussian noise i.e. regime 16 in Figure (3.14).

graph is similar to a number of others presented here in that the regression tree reported a number of spurious breaks when the series were short. Once the series became long enough, here about regime 16 (i.e. 256 data points per regime), this tendency disappeared.

Figure (3.15) presents the locations of the breaks for the regime 16 series as this was the shortest series for which ART did not overfit. As can be seen the tree reported the location of the two breaks as being near data points 256 and 512 respectively. As the break size decreased the breadth of an interval increased but there were no genuinely incorrect breaks reported.

## 3.3 Noisy Square Wave Simulations

The model used in these simulations was

$$y_t = \mu_{r_i} + \epsilon_t \tag{3.11}$$

where

$\mu_{r_i}$ = the mean of regime $r_i; i = 1, \ldots, 5$

$\epsilon_t$ = noise terms drawn from an N(0,1), gamma, or geometric distribution.

In all simulations $\mu_{r_i} = 0$ for $i = 1, 3, 5$ and $\mu_{r_4} = -\mu_{r_2}$. The value of $\mu_{r_2}$ was started at 2 standard deviations and was decremented to 0.05 in steps of 0.05. When the BP was used to detect breaks in the series, because the amount of computation required, the value of $\mu_{r_2}$ was sometimes decremented to 0.1 in steps of 0.1.

In essence the resultant series are square waves with an amplitude of break size with Gaussian (or other) noise of constant variance imposed on them.

### 3.3.1 CUSUM Tests

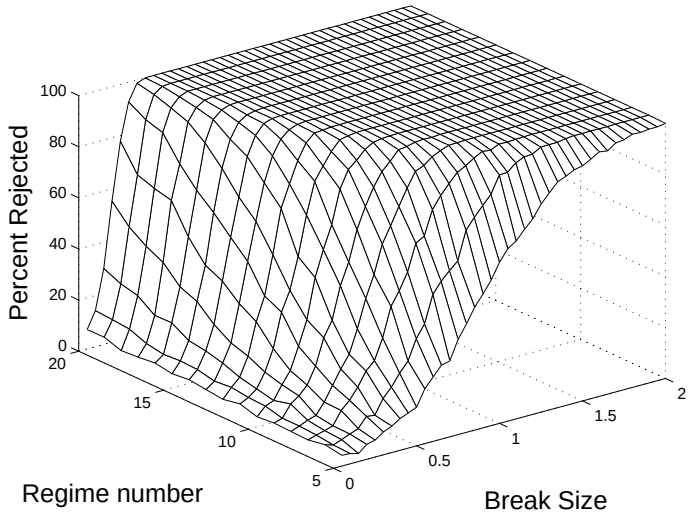


Figure 3.16: The percentage of noisy square wave series for which the null hypothesis was rejected at the five-percent level by a CUSUM test.

Details of the CUSUM test were outlined in Section (1.1.1). The purpose of the

simulations reported in this section was the same as outlined in Section (3.1.2) except they are for the noisy square wave.

The authors of the `strucchange` package, Zeileis et al. (2002), recommend doing a CUSUM test to check for the presence of a structural break before proceeding, if warranted, to run a full BP test on the series.

The results of the CUSUM tests are presented in Figure (3.16).

In the high plateau parameter region of the graph, tests for the location of a structural break should report at least one. As the break size and/or the regime length decreased the percentage of series for which a break ought to be detected declined. The rejection rate fell to a low of about five percent consistent with the chosen level of significance. It should be noted that the smallest break was 0.05 standard deviations. CUSUM was not applied to series with no breaks.

Some useful comparisons can been made between Figure (3.16) and Figures (3.17) and (3.20). Caution needs to be exercised to not over interpret the evidence. If the CUSUM test rejected the null hypothesis it only means at least one break should be reported. Both ART and the BP reported breaks in the parameter regions the CUSUM tests suggested they should. However, both ART and the BP had small areas where the reporting rate was zero, so performed marginally below expectations.

### 3.3.2 Gaussian Noise and Cost-Complexity Pruning

These simulations used the noisy square wave with Gaussian noise, no serial correlation and default cost-complexity pruning. The results from these simulations are presented in Figure (3.17).

The results showed mostly what could be expected from a statistical modeling procedure based on ordinary least squares. The problem with short series lay in reporting many spurious breaks.

There were some clear weaknesses with ART if the tree was pruned with the default cost-complexity pruning method.

When the break size was large and the regime size was long ART consistently

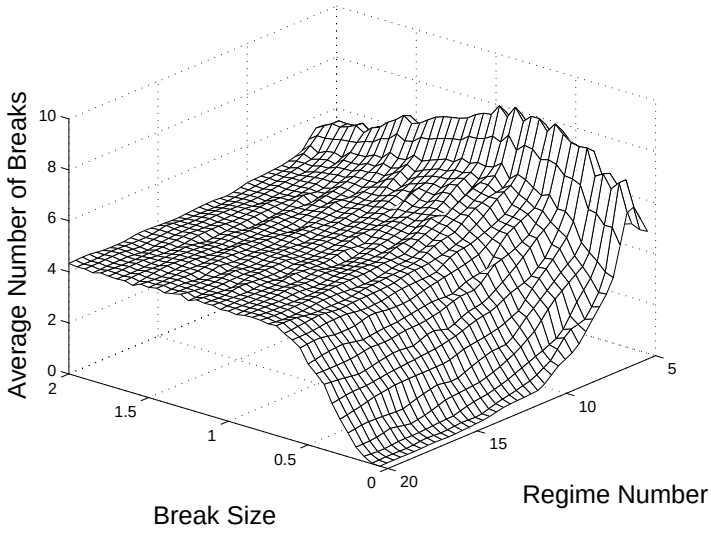


Figure 3.17: Average total number of breaks reported by ART in the noisy square wave simulations with Gaussian noise and cost-complexity pruning.

reported close to correct number of real breaks and generated few spurious candidate breaks without needing to prune further than that already done by cost-complexity pruning. The problem of slight over-fitting in the long regimes and large breaks sizes will be addressed in Chapter (4).

This is consistent with the previous simulations which showed a weakness in ART with cost-complexity pruning. With short series ART reported unacceptably high numbers of spurious breaks. This is addressed further in Section (3.3.3) which considers alternative pruning criteria and Section (4.2) which addresses the problem of an incorrect split in the root node.

### 3.3.3 Pruning Criteria

In this section we consider methods other than cost-complexity of selecting a best tree.

Esposito et al. (1997) evaluated a number of pruning procedures in the context of classification trees. In all but two of the data sets they examined, the one standard

error cost-complexity pruning procedure, outlined in Section (1.2.5), over-pruned the final tree. Su et al. (2004) evaluated a number of pruning criteria and reported the cost-complexity pruning developed by Breiman et al. (1993) was relatively insensitive in picking a best tree. See, for example, Figure (2) of Su et al. (2004). Their work suggested the BIC or RIC were excellent choices of tree selection criteria.

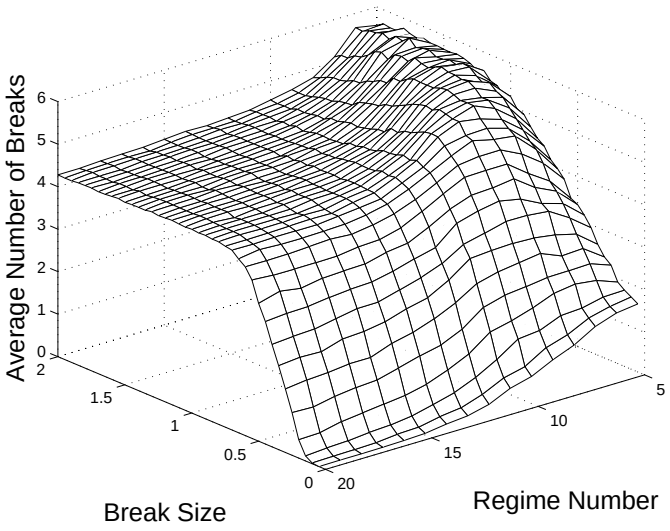


Figure 3.18: The average total number of breaks reported by ART as a function of regime length and break size with Gaussian noise and BIC tree selection.

Cross-validation is a standard procedure in recursive partitioning algorithms. According to Hastie et al. (2001, p221) it is a method of choice for tree procedures. As discussed earlier in Section (1.2.5) when applying cross-validation to time series there are some special problems to consider.

In light of the findings of Esposito et al. (1997) and Su et al. (2004) we evaluated the use of BIC pruning and leave-one-out cross-validation. Stone (1977) was able to show that the AIC and leave-one-out cross-validation were asymptotically equivalent. However, in finite samples they perform quite differently.

The series were simulated in the same way as in the previous section. The results for the simulations which used BIC tree selection are presented in Figure (3.18) and

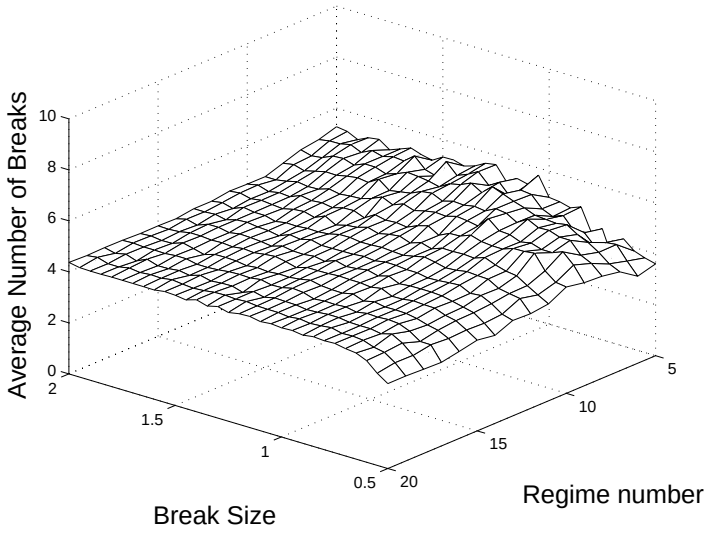


Figure 3.19: The average number of breaks reported by ART using leave-one-out cross-validation. Minimum break size is 0.5.

leave-one-out cross-validation in Figure (3.19).

If we compare Figure (3.17) with Figures (3.18) and (3.19) we see that with BIC tree selection ART pruned away many breaks in the short series resulting in a considerably lower "ridge" feature in those parameter regions. The leave-one-out cross-validation selected even smaller trees in the short series than did the BIC. The high sharp "ridge" of Figure (3.17) was reduced to a small rise in Figure (3.19).

As observed earlier, no additional pruning was required for long series. The broad flat area in the parameter regions on the left of Figures (3.17), (3.18) and (3.19) showed that in long series the BIC and leave-one-out cross-validation did not prune additional nodes from the trees.

Leave-one-out cross-validation was computationally expensive. To compare it with the BIC, the simulations for series with a regime length of 400 took 90.06 hours of CPU time for the leave-one-out cross-validations on a SunBlade 1000 with 750Mhz UltraSPARC-III CPU whereas they took 50.79 minutes for the BIC.

As the BIC took very little additional computational resources compared with leave-

one-out cross-validation it seems a good choice for routine tree selection criteria within ART. If the series was short, leave-one-out cross-validation may be worthwhile. ART with leave-one-out cross-validation was still faster than the BP.

### 3.3.4 Bai and Perron Procedure

For comparison purposes a number of simulations were run using the procedure of Bai and Perron (1998, 2003). Because of the considerable computational resources required only simulations up to regime length 256 (regime 16) were run, in the longer series the step size was 0.1 rather than 0.05 in the ART simulations, and as few as 30 replications were carried out. Up to four CPUs were simultaneously working on these simulations presented in this subsection ($2 \times 750$Mhz UltraSPARC III and $2 \times 1.5$Ghz UltraSPARC IIIi) over a period of approximately seven weeks.

The results are presented in Figure (3.20).

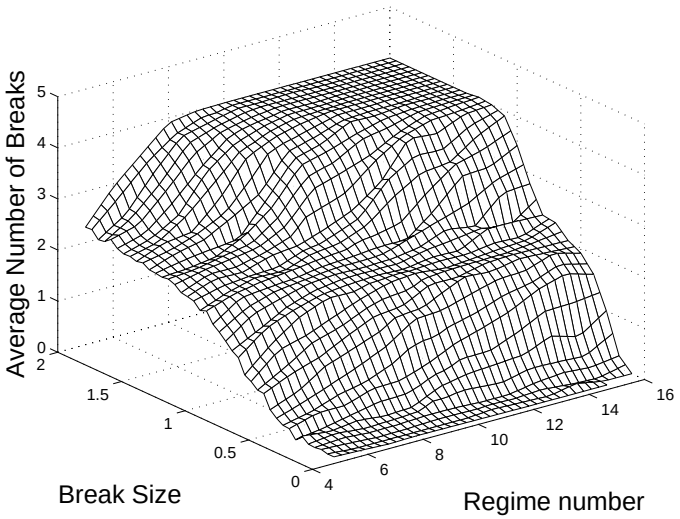


Figure 3.20: Average number of breaks reported by the BP as a function of regime length and break size.

The BP never reported more than four candidate breaks (see Figure 3.20), the

number of true breaks in the simulated series. The BP reported a confidence interval for each breakpoint. Thus it was meaningful to split the total number of breaks into correctly located breaks and spurious breaks. When the total breaks was split into correct and spurious breaks on the basis that the true break lay within or outside the reported confidence interval respectively, there were some series in which the BP appeared to have picked a wrong break. We investigated this by examining the raw data generated by the simulations. For the breakpoints examined individually the reported spurious breaks were the result of the true break usually laying no more than a few data points outside the reported confidence interval.

The histogram in Figure (3.21) shows the location of the candidate breaks in a parameter region where BP could be expected to have difficulty in locating the correct break. The regime length was 256 data points (i.e. regime 16) and the break size was 0.5 standard deviations. While there were breaks away from the actual locations at data points 256, 512, 768, and 1024, there was nothing like the truly spurious breaks in Figure (4.2) where ART split the root node incorrectly.

When reporting breaks in the series, the performance of ART and the BP appeared similar in the long series. In the short series the BP tended to under report breaks while ART over reported breaks. As we have seen, most of the over reporting by ART could be corrected by BIC pruning or leave-one-out cross-validation.

One other feature of the data which should be remarked upon is the "bench" feature visible in Figures (3.20), and (3.24). It appeared to be real as it was seen in other simulations which have been omitted for reasons of space. We have no explanation for its presence.

### 3.3.5 Gamma and Geometric Error Structure

We ran simulations with both gamma and geometric error structures. Similar to the results presented in Section (3.1.4) the results for the gamma and geometric errors were almost indistinguishable and we omit the results for the geometric errors for reasons of space, having presented results for geometric errors in Section (3.1.4).

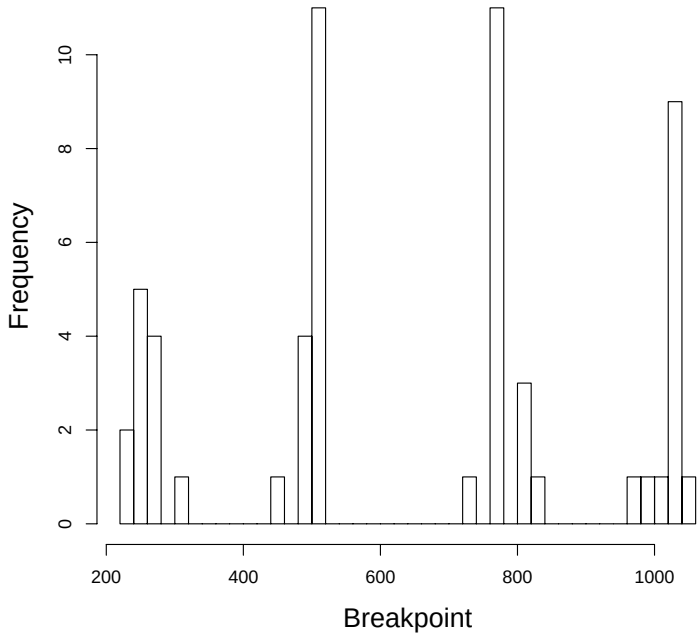


Figure 3.21: Location of candidate breaks reported by the BP for regime length 256 (i.e. regime 16) and break size 0.5 standard deviations.

The gamma distribution was outlined in Section (3.1.4) and Equation (3.4). Recall that we used $\mu = \sigma^2 = 2$ in the simulations.

Figure (3.22) shows one realization of a series with regime length 400 and break size two. The skewed nature of the noise can easily be seen. The simulations were as outlined for the noisy square wave.

The results are presented in Figure (3.23) for ART and Figure (3.24) for the BP.

As stated above ART's performance in locating the breaks was almost indistinguishable from the series with geometrically distributed noise. The "ridge" feature was slightly lower in Figure (3.23) than in the geometric noise case. The broad flat area of Figure (3.23) started at approximately break size 0.85 whereas it started at approximately break size 0.95 in the geometric noise. Overall, ART handled gamma noise

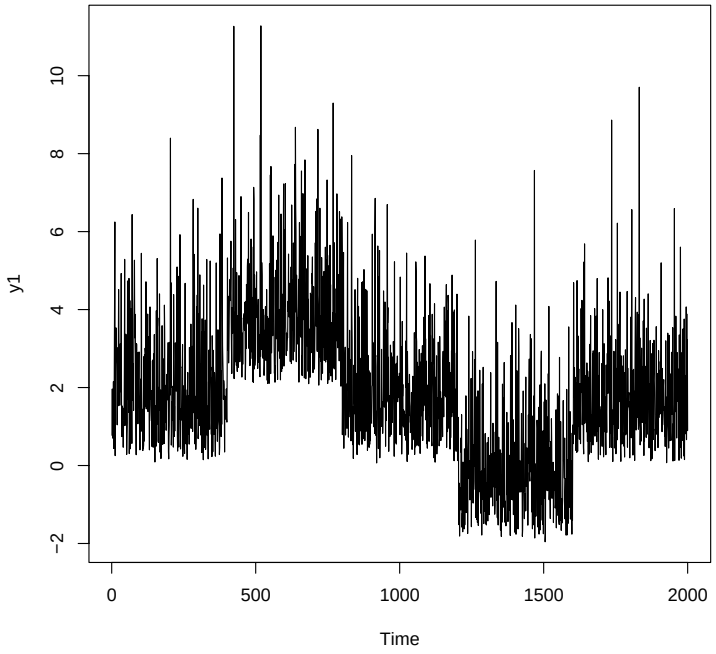


Figure 3.22: One realization of a square wave with gamma distributed noise with $\alpha = 2, \beta = 1$.

slightly better than geometric noise.

As seen in the other simulations the BP tended to miss breaks rather than report any truly spurious breaks. In these particular combinations of parameters the BP never consistently reported four breaks (Figure 3.24). Also, as remarked upon earlier, its ability to report breaks as the break size decreased seemed to decline in two stages for reasons that are not clear.

## 3.4 Series with Heteroscedasticity

We examined ART's robustness to heteroscedasticity by simulating series with a break at the mid-point and different standard deviations in the two halves. The first half

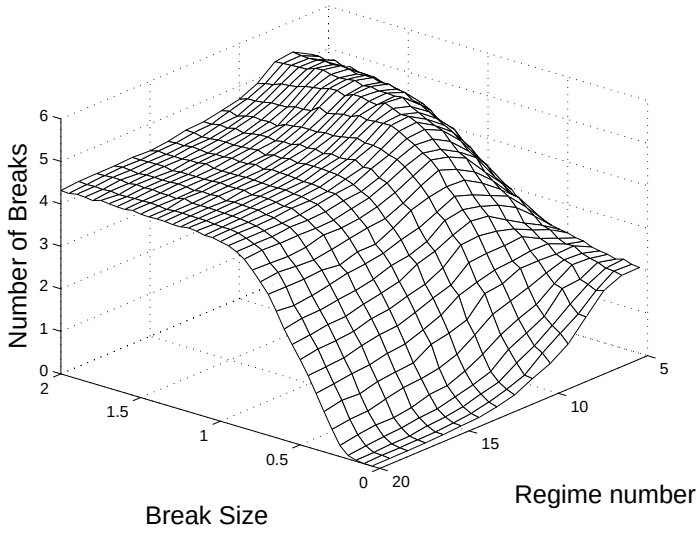


Figure 3.23: Average number of candidate breaks reported by ART for a series as a function of regime length and break size with gamma distributed noise when $\alpha = 2, \beta = 1$ and using BIC tree selection.

always had a standard deviation of one. The break size is stated in standard deviations of the first half. The second half had a standard deviation ranging from 1 to 2.95. We examined two lengths of series, 800 and 1800 data points. We did not run BP for comparison due to the excessive computational times it would require.

The results for the 1800 data point series are presented in Figure (3.25) the remainder are omitted for reasons of space. ART was more robust to heteroscedasticity in the longer series than in the shorter series. This was consistent with the other observations presented in this chapter that the problem of over-fitting declined with increasing series length.

## 3.5 Staircase Simulations

This set of simulations was run to compare the effectiveness of leave-one-out cross-validation with cost-complexity and the BIC as a pruning techniques in series with

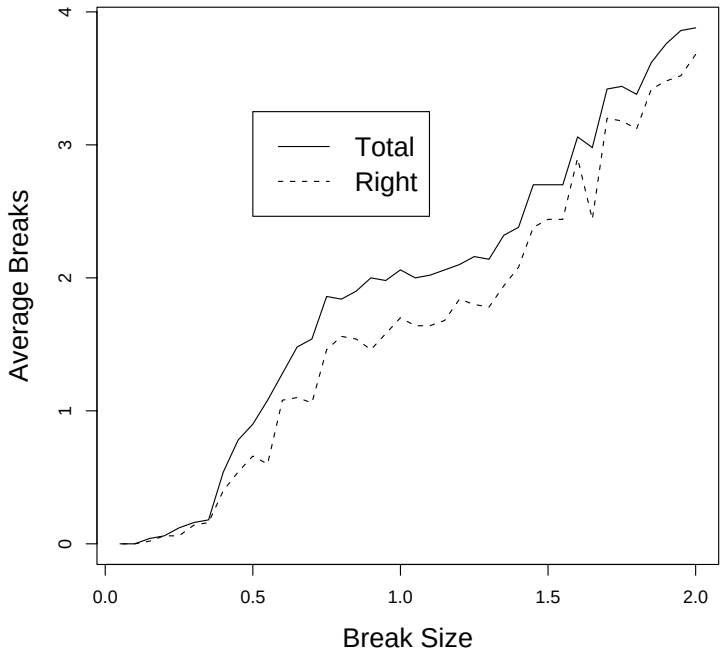


Figure 3.24: Average number of candidate and correctly reported breaks by the BP for a series with regime length 144 data points (i.e. regime number 12) as a function of break size with gamma distributed noise and $\alpha = 2, \beta = 1$.

varying levels of serial correlation. The model was a five regime series with

$$\mu_{(i+1)} = i \times \text{break size}; i = 0, 1, 2, 3, 4$$

with standard normal noise as the error term and regime length 200.

The models are in essence a noisy, correlated staircase as can be seen in Figure (3.26).

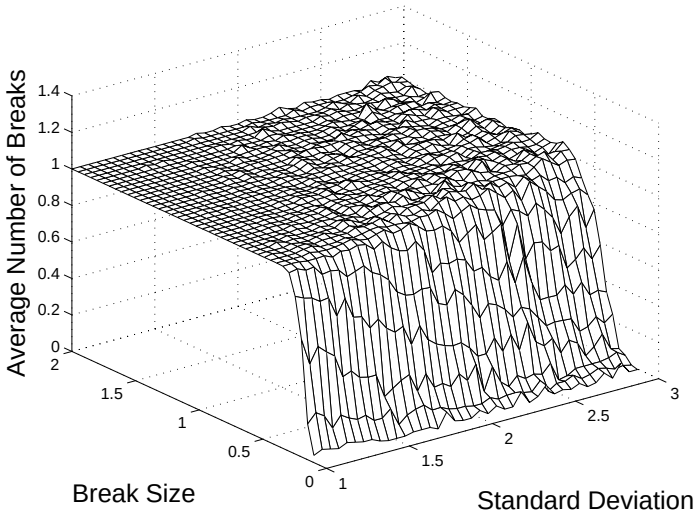


Figure 3.25: Average total number of breaks reported by ART in 1800 data point series with heteroscedasticity and cost-complexity pruning. The series had one true break.

The three models used were as follows.

$$AR(1) \quad \phi_1 = 0.4$$

$$AR(2) \quad \begin{cases} \phi_1 = 0.4 \\ \phi_2 = 0.24 \end{cases}$$

$$AR(5) \quad \begin{cases} \phi_1 = 0.4 \\ \phi_2 = 0.24 \\ \phi_3 = 0.144 \\ \phi_4 = 0.090 \\ \phi_5 = 0.054 \end{cases}$$

The results are presented in Figure (3.27). A single line in the graph presents the results for each of the three tree selection methods as they were close to indistinguishable and could only be resolved at much higher magnifications.

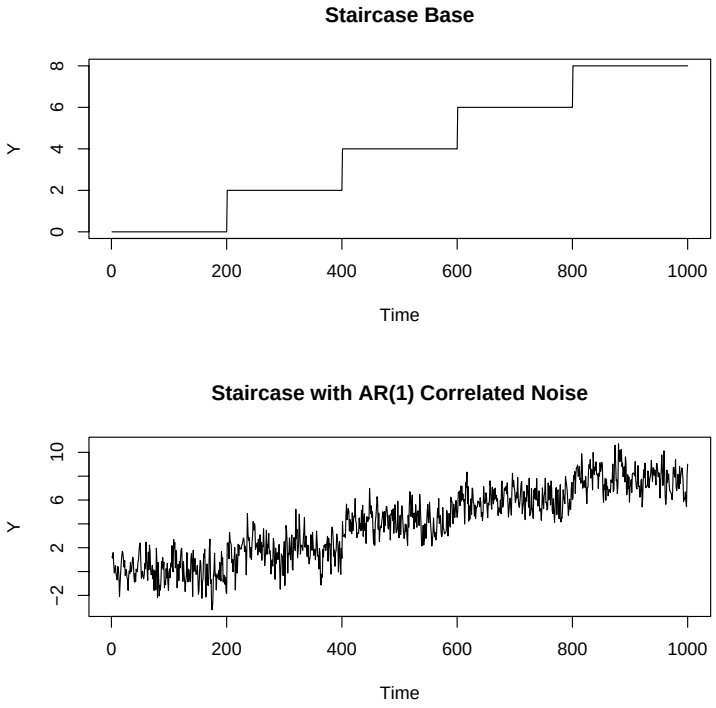


Figure 3.26: The noisy staircase. Top panel is the base staircase, lower panel is the staircase with AR(1) correlated noise.

In the AR(1) case leave-one-out cross-validation did prune a few more spurious breaks than the other two techniques but these were so few the improvement could not be seen graphically.

In the case of AR(2) correlations leave-one-out cross-validation produced slightly lower total reported breaks when the break size was small whereas the cost-complexity and BIC tree selection methods produced identical results.

In the AR(5) correlation the leave-one-out cross-validation again produced only slightly lower total breaks when the break size was less than approximately 1.1 standard deviations compared to the other two methods.

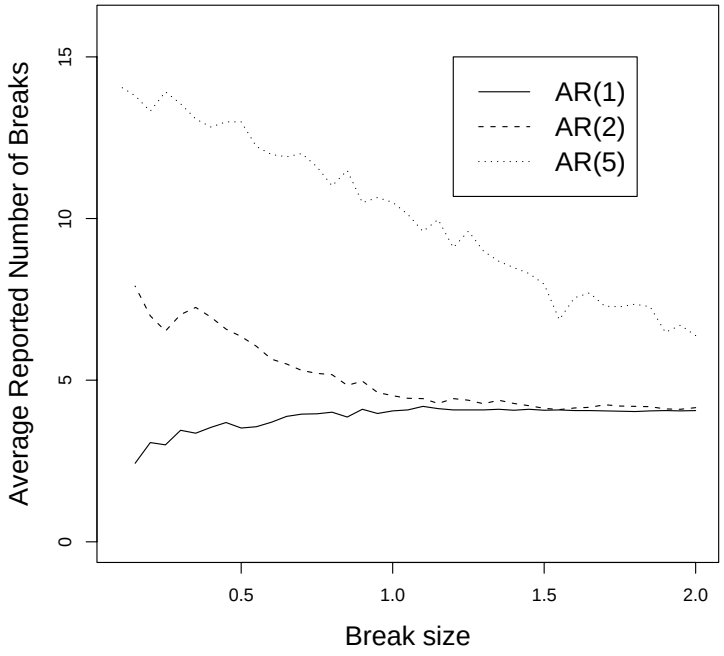


Figure 3.27: Average number of breaks reported by ART in the noisy staircase simulations with AR(1), AR(2), and AR(5) serial correlation for cost-complexity pruning, BIC pruning and leave-one-out cross-validation. See the text for more details.

## 3.6 Robustness to Outliers

A set of four simulations were run to assess ART's robustness to outliers. A series of 400 data points was generated from standard normal noise and tested to ensure ART did not report a break in the series. The data point at locations 20, 40, 60, and 80 were then gradually moved up until ART reported a break. The distance of the outlier from the mean was recorded in standard deviations.

The results are presented in Figure (3.28). As can be seen, the closer the outlier was to end of a series the easier it was to induce ART to report a spurious break. A small number of trees were visually examined and in most cases ART's solution to the problem of an outlier was to isolate it into a short segment of `mincut` in length.

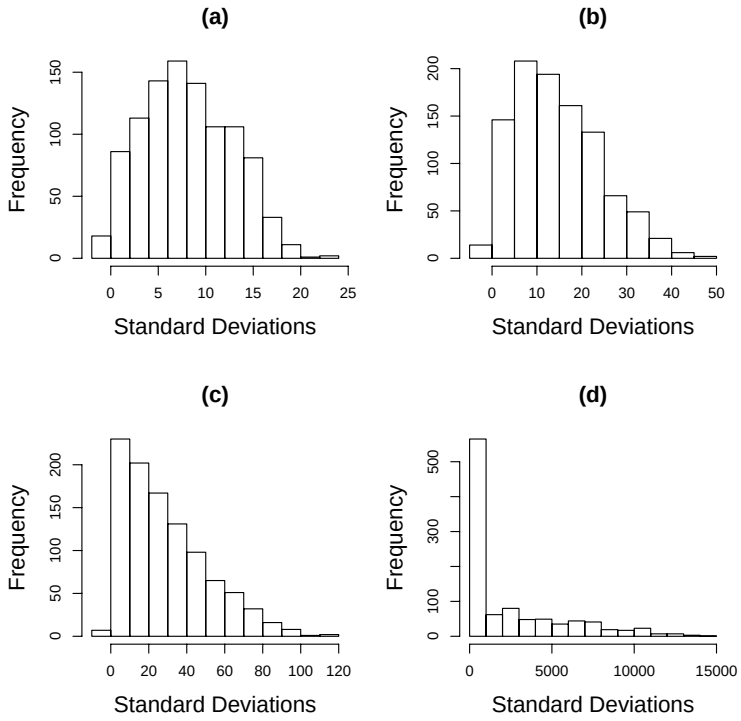


Figure 3.28: Histograms of the size of an outlier required to induce ART to report a break in a series of length 400 at four different positions. (a) Data point 20. (b) Data point 40. (c) Data point 60. (d) Data point 80.

## 3.7 Confidence Intervals

Regression trees are a non-parametric method, thus it is not straight forward to obtain confidence intervals around the breakpoints. The usual way to obtain confidence intervals is to use bootstrap methods. The bootstrap is due to Efron (1979). Efron built on past work in developing the bootstrap by drawing on three papers by Hartigan (1969, 1971, 1975), and one each by Jaeckel (1972) and Maritz and Jarrett (1978). It has become a standard tool in non-parametric statistics. For time series the presence of serial correlations in the data makes bootstrapping difficult but this has been addressed by the sieve bootstrap discussed by Alonso et al. (2002) and Park (2002). It is impor-

tant to recall that regression trees fit piece-wise constant functions to the data without taking account of any serial correlation. Thus only in very limited circumstances can we meaningfully use the bootstrap.

### 3.7.1 Method

To bootstrap a series we first fit and remove a model. Because ART fits piecewise constant functions to the data the means of these constant functions are subtracted from the data leaving, one hopes, *iid* residuals. The residuals are sampled with replacement using a uniform distribution as the sampling distribution to create a noise series of the same length as the original. Finally, the means are added to the noise to create a set of bootstrapped data.

We used 1000 replications of the bootstrapped series to determine the confidence interval. Because ART sometimes missed breaks and sometimes reported spurious breaks care was required in interpreting the results from the bootstrapping. The process we used to determine the confidence interval and whether the break was real or spurious can be illustrated by an example.

Consider a series with two breaks with $1 < b_1 < b_2 < T$ where $T$ is the length of the series, and $b_1$ and $b_2$ are the breakpoints in the original series. Below, $b_c$ is a candidate breakpoint from the bootstrapped series. The following criteria were applied to the $b_c$ from the bootstrapped series.

1. If the candidate breakpoint was between 1 and $\lfloor b_1/2 \rfloor$ or between $\lfloor (b_2+t)/2 \rfloor + 1$ and $T$ it was regarded as spurious where $\lfloor b_1/2 \rfloor$ denotes the highest integer less than $b_1/2$.

2. If there was a single candidate breakpoint $(\lfloor b_1/2 \rfloor + 1) \leq b_c \leq \lfloor (b_1 + b_2)/2 \rfloor$ it was used in the calculation of the confidence interval for $b_1$. A similarly defined interval was used for each breakpoint.

3. If there was no candidate breakpoint in the interval the breakpoint was counted as missed.

4. If there was more than one candidate breakpoint in the interval the breakpoint closest to $b_1$ (or $b_2$) was selected for the calculation of the confidence interval, the other(s) was counted as spurious.

### 3.7.2 Single Break Uncorrelated Series

In this section we examine the confidence intervals for series with uncorrelated observations and a single break. We generated series with Gaussian, gamma and geometric distributed noise. The gamma and geometric distribution parameters were as in Section (3.1.4).

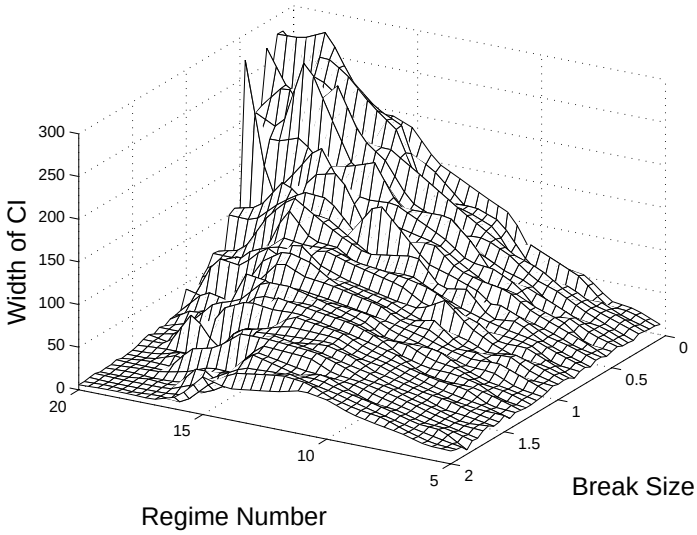


Figure 3.29: 95% confidence interval widths for series with a single break, uncorrelated observations and Gaussian noise.

Figure (3.29) presents the results for series with Gaussian Noise. Figure (3.30) presents the results for series with gamma distributed noise. The results for the geometrically distributed noise were almost indistinguishable from the gamma noise and are omitted for reasons of space. For comparison purposes Figure (3.31) presents the confidence intervals for the BP for series with Gaussian noise where these could be obtained.


Figure 3.30: 95% confidence interval widths for series with a single break, uncorrelated observations and Gamma noise.

As the confidence interval was assumed to be symmetric about the breakpoint in these series we report the confidence interval widths. While both graphs from the bootstrapping share a similar overall shape the widths of the gamma (and geometric) noise confidence intervals were wider than their Gaussian counterparts. The method we have used to determine whether a break is real or spurious restricted the confidence interval to a maximum of the length of a regime. Thus the apparently good performance in the shorter regimes (see the parameter regions on the left hand side of Figures 3.29 and 3.30) was simply an artifact of the confidence interval width being constrained by the regime length.

Even for moderate break sizes the confidence intervals were surprisingly wide. This was not an artifact of the bootstrap process as the BP had similarly wide confidence intervals where these could be determined. For regime 20 (regime length 400), break size 0.85 standard deviations and Gaussian noise the bootstrap reported a confidence interval width of 46 data points. The BP reported a confidence interval width of 36 data points for the same parameters.

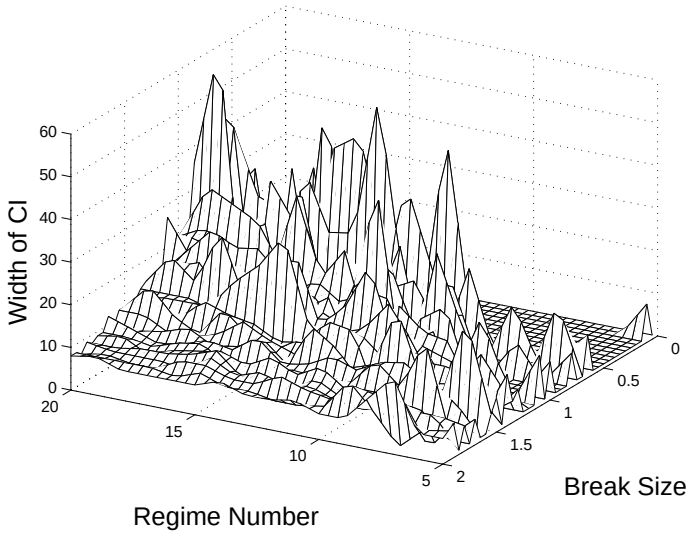


Figure 3.31: 95% confidence interval widths reported by the BP for series with a single break at the mid-point, uncorrelated observations and Gaussian noise.

### 3.7.3 Noisy Square Wave Bootstrapping

In Chapter (4) we give details of the incorrect splitting of the root node when ART was applied to the noisy square wave. Figure (3.32) presents the results of bootstrapping the confidence interval for the second break in noisy square series with regime length 400 data points. In the noisy square wave simulations there was a tendency for ART to displace the break away from the true break point at 800. For these series an asymmetrical confidence interval could be expected and indeed was reported. The BP was not run on series of this length because of the long compute times required. However, it was run on series with regime length 256 data points. In these series the BP did not report any systematic departure from a symmetric confidence interval. Thus the asymmetrical confidence interval must be attributed to the *greedy* algorithm[2] which ART uses.

An alternative is to consider whether bootstrapping can be used to distinguish be-

[2]The term *greedy* algoritm was difined in Section (1.2.4)

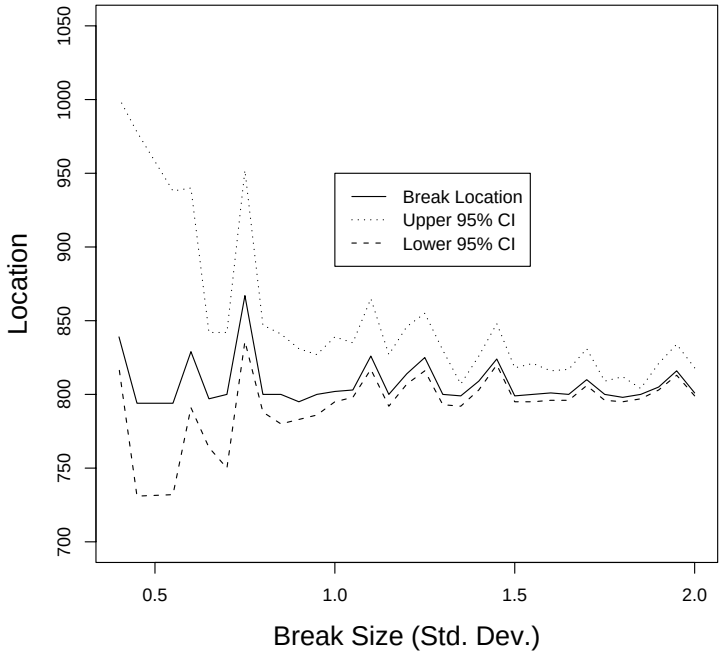


Figure 3.32: 95% confidence interval widths for the second break of the noisy square wave, uncorrelated observations and Gaussian noise.

tween true and spurious breaks. In the noisy square wave series there are two categories of spurious breaks to consider.

1. Incorrect splitting of the root node owing to the failure of the *greedy* algorithm.

2. Other breaks arising from random fluctuations.

We discuss the first of these types of breaks here and defer the second to Section (3.7.4).

We suggest in Section (3.7.4) below that if the bootstrap missed a break more than 12 percent of the time then it was likely to be spurious. We applied that criteria to 100 noisy square wave series with regime length 400, break size two standard deviations and which incorrectly split the root node. The bootstrap correctly flagged all 100 root node

breaks as spurious. However, it also flagged a further 21 breaks as spurious but which were real. A visual examination of the results showed that this occurred when the real and spurious breaks were very close to each other. Given the algorithm presented in Section (3.7.1), it was not possible to generate overlapping confidence intervals. However, one can estimate the "natural" confidence interval for these breakpoints from the series which did not split the root node incorrectly. In the cases in which both the real and spurious breakpoints were flagged as spurious the "natural" confidence interval would either have overlapped both breakpoints or nearly so.

### 3.7.4 Detecting Spurious Breaks

Throughout the work presented in this chapter we often observed that ART reported spurious breakpoints. For series with serial correlations these reported breakpoints could often be attributed to long excursions away from the series mean. For series without serial correlations breakpoints were reported which were just the result of random fluctuations. Several methods were tried to reduce the number of spurious breaks, these were detailed in Sections (3.1.7) and (3.3.3). Each produced some improvement but none were entirely successful. In the case of random fluctuations it seemed reasonable to suspect that real and spurious breaks could be distinguished through the use of the bootstrap. This is investigated in this section.

As the problem of spurious breaks was more serious in short series we selected series of 200 data points for our investigations.

We generated series without any real breaks and selected for bootstrapping 1000 for which ART had reported one or more spurious breakpoints. These 1000 series yielded 3251 spurious break points. These series were bootstrapped 1000 times and the number of times the reported breakpoint was missed in the bootstrapped series was recorded. The results are presented in Figure (3.33).

As can be seen the distribution had a peak in the bar representing 101 to 150 misses per thousand bootstrapped series. This should be compared with Figure (3.34) which is a histogram of the number of times a real break of size 0.1 standard deviations was

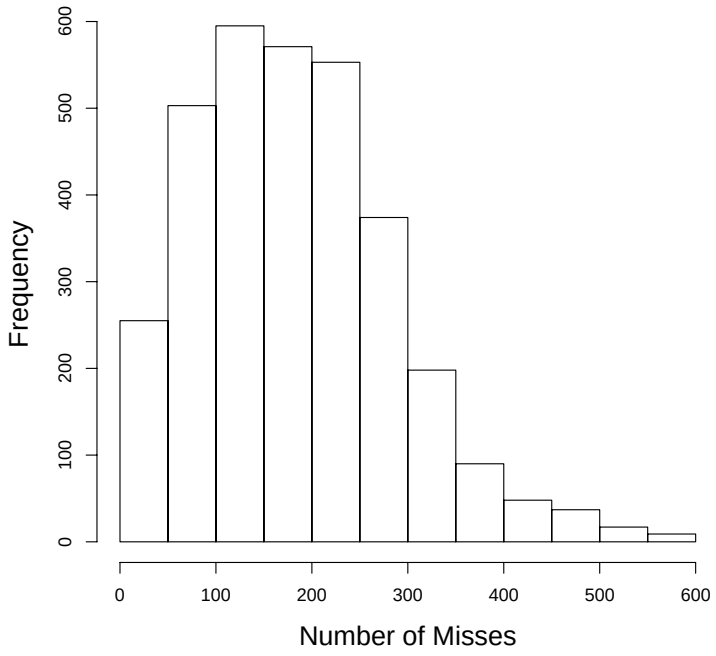


Figure 3.33: Histogram of the number of times a spurious break in a series of 200 *iid* Gaussian data points was not reported in a bootstrapped series.

not reported in a bootstrapped series. Even with such a small break the number of misses was usually less than 100 per 1000 bootstrapped series. Thus the histograms show qualitatively different behaviours for real and spurious breaks. This suggested that if a reported breakpoint was missed in more than 12 percent of the bootstrapped series we can be at least 95% confident that the break was spurious.

## 3.8 Computation Times

We ran some comparisons of compute times between ART and the BP. These are presented in Table (3.2). As can be seen the compute times for the BP increased rapidly with series length while the compute times for ART were short and increased

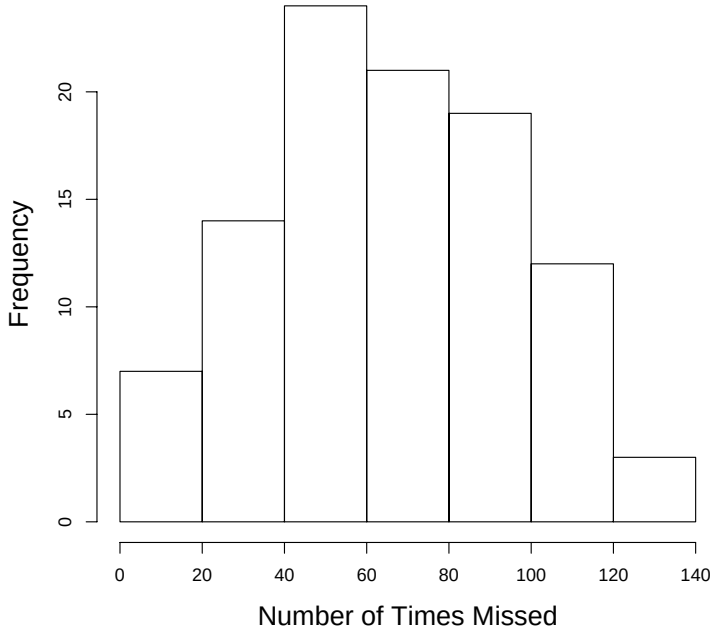


Figure 3.34: Histogram of the number of times a real break of size 0.1 standard deviations in a series of 200 data points with Gaussian noise was not reported in a bootstrapped series.

only slowly. Given the disparity between the compute times, ART is a useful addition to the practitioner's toolbox when analyzing long time series.

## 3.9 Conclusions

We are now able to answer the six questions posed at the beginning of this chapter.

1. In the absence of serial correlations, ART did impose spurious breaks when the series was short but this tendency disappeared as the series became longer. This was seen in single, multiple break and heteroskedastic simulations.

2. ART was robust to negative serial correlation and a small amount of positive

Table 3.2: Comparison of processor times (hours:minutes:seconds) required to run the BP with minimum segment length set to 0.05 and ART for series of various lengths. The times for ART were so short that it was difficult to get accurate timings. All times were run on a SunBlade 1000 with 750Mhz UltraSPARC-III CPU and 2Gb RAM.

| Series Length | Bai & Perron | ART |
|---|---|---|
| 500 | 0:03:10.05 | 0:00:00.03 |
| 1000 | 0:14:08.24 | 0:00:00.04 |
| 1500 | 2:04:13.25 | 0:00:00.05 |
| 2000 | 3:41:56.75 | 0:00:00.07 |
| 2500 | 6:18:34.35 | 0:00:00.08 |

correlation, but in this regard ART was no worse than the BP.

3. Leave-one-out cross-validation can be used for tree selection but was computationally expensive.

4. ART did report higher numbers of breaks for series with non-Gaussian noise structures but not excessively so.

5. ART's robustness to outliers depended on the outlier location. For an outlier 20 data points from the end of the series a outlier needed to be a mean of over 8 standard deviations away from the series mean to cause ART to spuriously report a break when none existed. This increased to a mean of over 2300 standard deviations when the outlier was at data point 80.

6. It is possible to obtain a confidence interval for the breakpoints using the bootstrap. However, because regression trees only fit piece-wise constant functions to the data, this is currently limited to series without serial correlations and may thus be of little practical use.

In addition to answering these six questions a number of other observations can be made.

1. For routine tree selection the BIC was found to be a good choice.

2. For series with less than approximately 500 data points leave-one-out cross-validation yielded a worthwhile reduction in spurious breaks reported. However, for series of this length and shorter, if time permits, a CUSUM test followed by a BP, if appropriate, seemed a more reliable method of locating structural breaks.

3. In the absence of serial correlation the break size is the dominant factor in ART's ability to correctly locate a break.

4. ART showed greater robustness to a non-normal noise structure than did the BP.

5. Plotting the tree gives a visual representation of the breakpoints and the mean values of the regimes reported. See the trees presented in Section (1.2) and the case study in Chapter (5).

6. In the simulations presented in this chapter ART was found to be computationally faster than both CUSUM and the BP. This makes it especially suited to the analysis of long time series.

7. ART is a useful addition to current tests for the detection and location of structural breaks in the mean of a time series. It is particularly useful for series with many data points for which computational times for the BP are excessive.

## 3.10 Publications and Presentations

Material from this chapter is contained in Rea et al. (2006), Cappelli et al. (2007), and Cappelli et al. (2008). The poster paper on which Rea et al. (2006) was based placed second in the poster paper competition at the International Workshop on Statistical Modeling held at Galway, Ireland, in July, 2006.

# Chapter 4

# Enhanced Temporal Pruning

## 4.1 Introduction

In the simulations we noticed that there were often spurious breaks even in the simulations with long regimes and large break sizes. For example, in Figure (3.19) the plateau region was above four with the number of spurious breaks settling at an average of about half a break per series. In this chapter we present the evidence for a systematic generation of spurious breaks by ART, detail the reasons for these breaks including some theoretical analysis, propose a method of solution, and evaluate its effectiveness.

## 4.2 Too Many Terminal Nodes

In the square wave simulations presented in Section (3.3) there were a high number of trees with greater than five terminal nodes. From the simulations with regime length 100 (regime number 10), break size of two standard deviations and which used cost-complexity pruning, 100 trees with more than five terminal nodes were selected. Of these, 89 had an incorrect split in the root node and 20 had at least one incorrect split at other nodes in the tree. These two numbers add to more than 100 because some trees had seven or more terminal nodes.

| Number of Breaks | Count | Percentage |
|---|---|---|
| Three | 0 | 0.00 |
| Four | 5504 | 55.04 |
| Five | 4292 | 42.92 |
| Six | 196 | 1.96 |
| Seven | 8 | 0.08 |
| Eight | 0 | 0.00 |

Table 4.1: Comparison of number of breaks reported in trees grown on simulations with regime length of 100 and break size of two standard deviations.

Cost-complexity pruning was not the only way to automatically prune a regression tree. The use of an information criteria such as the BIC (Schwarz, 1978) or leave-one-out cross-validation, under certain circumstances, improved the performance of ART. See Section (3.3.3) for more details. These were shown to be useful in pruning spurious breaks from the maximal tree.

By far the more serious problem was the high rate of incorrect splitting of the root node. Table (4.1) gives the proportions of trees with various numbers of breaks. In all cases examined, the five, six, and seven break trees had an incorrect split in the root node. In the five break trees this was the only error in the tree.

In Figures (4.1) and (4.2) we present the distributions of the breakpoints for 10,000 trees with four and five breaks respectively for regime length 100 and break size two standard deviations. The main observable difference between these two types of trees is the presence of many breaks in the third regime for trees with five breaks. We further examined the the distribution of the splits in the root node for 1,000 trees with regime length 144 and break size two standard deviations. It was clear that the split of the root node was being made in the third regime, frequently well away from either of the true breakpoints.

ART, in its default tree growing and pruning scheme, had no way to recover from this problem. In examining the six and seven terminal node trees it was often obvious

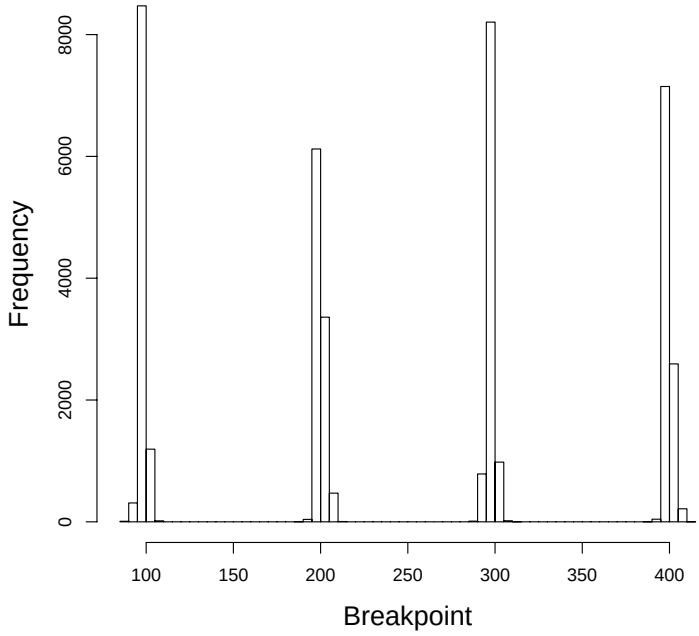


Figure 4.1: Distribution of breakpoints for trees with four breaks. Regime length was 100 and break size was two standard deviations.

that ART's solution was to isolate the incorrect break into a small terminal node, sometimes of minimum cut size.

Figure (4.3) is an example of a regression tree of a series for which the root node was split incorrectly. Given that the true breakpoints are at 144, 288, 432, and 576 the four split points on the lower non-terminal nodes were done almost perfectly.

The regression tree algorithm calculates a RSS for all allowable splits and chooses the minimum. Figure (4.4) presents the RSS for this series. ART makes one pass through this data and selects the break which gives the greatest reduction in RSS. The lowest RSS was at data point 306, hence it was chosen as the split point. The five regimes are clearly visible.

The third regime was different to the other four. It had a regime mean equal to the

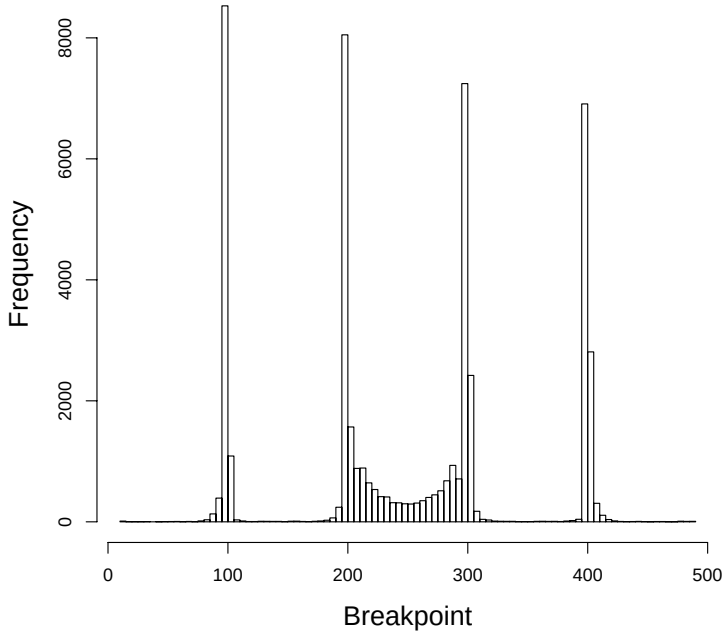


Figure 4.2: Distribution of breakpoints for trees with five breaks. Regime length was 100 and break size was two standard deviations.

series mean and passed through the mid-point. Under these conditions the RSS of the third regime without any serial correlation is a random walk with its two end points fixed by the RSS when the series is split at breakpoints two and three.

## 4.3 Regime Three

In this section we examine the theoretical RSS for the noisy square wave with Gaussian noise. By using these results and applying them to the third regime we present a way of obtaining the distribution of the split of the root node seen in Figure (4.2).

If $X_1, \ldots, X_n$ are independent normal random variables drawn from an $N(\mu_i, 1)$

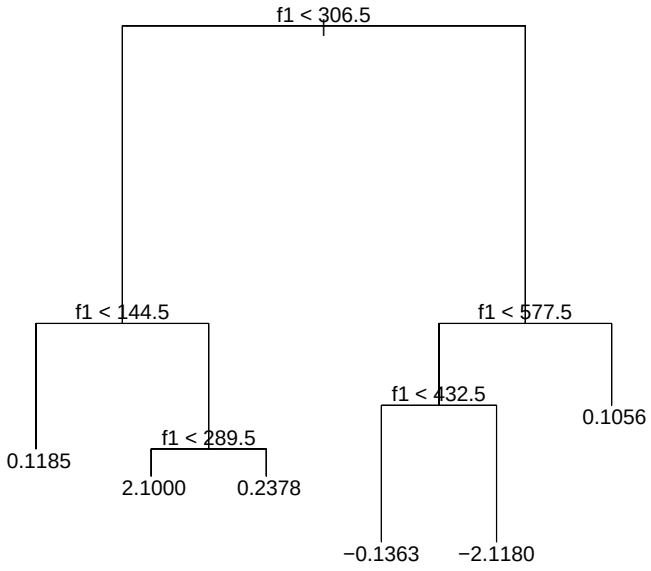


Figure 4.3: Regression tree for a series with an incorrectly split root node.

distribution, then the random variable

$$Y = \sum_{i=1}^{n} X_i^2$$

has a non-central $\chi^2$-distribution with n degrees of freedom and non-centrality parameter

$$\mathcal{N} = \frac{1}{2} \sum_{i=1}^{n} \mu_i^2. \tag{4.1}$$

It can be shown that the moment generating function (Blum and Rosenblatt, 1972, p254) for this distribution with $n$ degrees of freedom is given by

$$M_Y(t) = \frac{1}{(1-2t)^{\frac{n}{2}}} e^{\frac{2\mathcal{N}t}{1-2t}}. \tag{4.2}$$

We can calculate the mean and variance of this distribution from

$$\mu = \left. \frac{dM_Y(t)}{dt} \right|_{t=0} \tag{4.3}$$

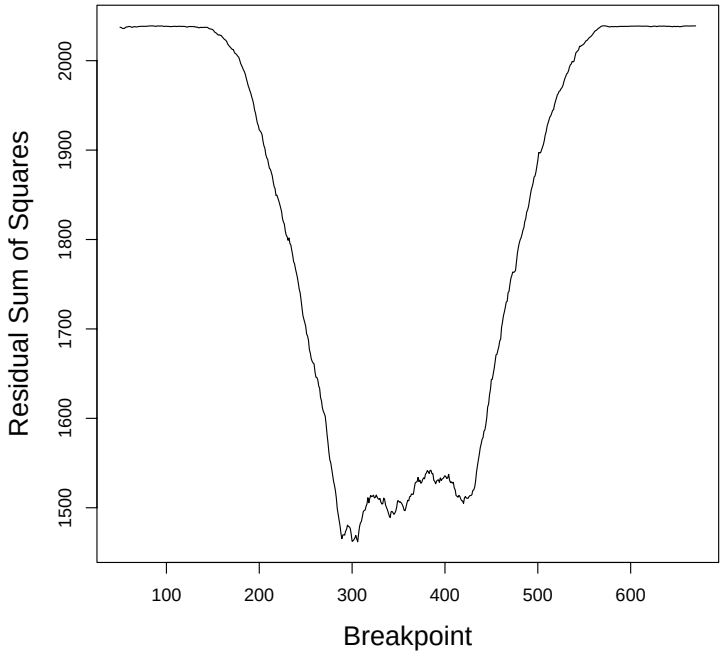


Figure 4.4: Residual sum of squares as a function of the breakpoint for the series whose regression tree is presented in Figure (4.3).

and

$$\sigma^2 = \left.\frac{d^2 M_Y(t)}{dt^2}\right|_{t=0} - \left(\left.\frac{dM_Y(t)}{dt}\right|_{t=0}\right)^2.$$

Differentiating (4.2) and simplifying we obtain

$$\frac{dM_Y(t)}{dt} = \left[\frac{n}{(1-2t)^{\frac{n+2}{2}}} + \frac{2\mathcal{N}}{(1-2t)^{\frac{n+4}{2}}}\right] e^{\frac{2\mathcal{N}t}{1-2t}}$$

and so

$$\left.\frac{dM_Y(t)}{dt}\right|_{t=0} = n + 2\mathcal{N} \tag{4.4}$$

a result which is intuitively reasonable.

Taking second derivatives and simplifying through much tedious algebra we obtain

$$\begin{aligned}\frac{d^2M_Y(t)}{dt^2} =& \frac{n^2e^{\frac{2\mathcal{N}t}{1-2t}}}{(1-2t)^{(\frac{n+4}{2})}} + \frac{2n\big(\frac{2\mathcal{N}}{1-2t} + \frac{2\mathcal{N}}{(1-2t)^2}\big)e^{\frac{2\mathcal{N}t}{1-2t}}}{(1-2t)^{(\frac{n+2}{2})}} \\ &+ \frac{2ne^{\frac{2\mathcal{N}t}{1-2t}}}{(1-2t)^{(\frac{n+4}{2})}} + \frac{\big(\frac{8\mathcal{N}}{(1-2t)^2} + \frac{16\mathcal{N}t}{(1-2t)^3}\big)e^{\frac{2\mathcal{N}t}{1-2t}}}{(1-2t)^{(\frac{n}{2})}} \\ &+ \frac{\big(\frac{2\mathcal{N}}{1-2t} + \frac{4\mathcal{N}t}{(1-2t)^{\frac{n}{2}}}\big)^2e^{\frac{2\mathcal{N}t}{1-2t}}}{(1-2t)^{(\frac{n}{2})}}.\end{aligned}$$

Then

$$\left.\frac{d^2M_Y(t)}{dt^2}\right|_{t=0} = n^2 + 4\mathcal{N}n + 2n + 8\mathcal{N} + 4\mathcal{N}^2.$$

Thus

$$\left.\frac{d^2M_Y(t)}{dt^2}\right|_{t=0} - \left(\left.\frac{dM_Y(t)}{dt}\right|_{t=0}\right)^2 = 2n + 8\mathcal{N}.$$

These results can be used to calculate the theoretical value for the residual sum of squares for the whole series and for the series when it is split into two pieces at breakpoint two. We do this for the noisy square wave with regime length 144 (regime 12 in the results plotted in Figure 3.17) with standard normal noise and break size two standard deviations. We obtain

$$E[RSS] = 1872.$$

This compares very favourably with the values $\mu_{rss} = 1871.4$ obtained from a thousand replications of the series.

Similarly we can calculate the theoretical residual sum of squares when the series is split at the second breakpoint.

$$E[RSS_{bp2}] = \sum_{i=1}^{720} E[(x_i - \mu_i)^2].$$

We obtain

$$E[RSS_{bp2}] = 1392.$$

This value compares very favourably with the value $\mu_{rss-bp2} = 1391$ obtained from the same thousand replications.

We can calculate a theoretical value for the variance of the distribution of the residual sum of squares when the series is split at breakpoint two by using a mixture of three non-central $\chi^2$-distributions. Following similar reasoning to above we obtain a value of $\sigma^2 = 4128$ or $\sigma = 64.25$. The empirical value as obtained from the simulations, namely $\sigma = 62.8$, compares very favourably with this.

A Brownian Bridge is often defined as

$$X(t) = W(t) - tW(1); 0 \leq t \leq 1$$

where $W(\cdot)$ is the Wiener process (Hoel et al., 1987, p123). This is easily generalized to

$$X(t) = W(t) - \left(\frac{t - t_1}{t_2 - t_1}\right)(W(t_1) - W(t_2))$$

where $W(t_1) = a$ and $W(t_2) = b$ with $a, b$ constants. We can adapt this to understand the properties of the third regime and how it gives rise to an incorrect split in the root node.

In the general case if the distribution of $W(t), t \in (t_1, t_2)$ is standard normal, as they are in the simulations, then

$$\mu_t = a + \left(\frac{t - t_1}{t_2 - t_1}\right)(b - a)$$
$$\sigma_t^2 = \frac{(t - t_1)(t_2 - t)}{t_2 - t_1}$$

The expected values of $a$ and $(b-a)$ are 1392 and 64.25 respectively as obtained above. Without loss of generality this can be simplified to $a = 0$ and $b = 64.25$.

Figure (4.5) shows three confidence interval envelopes for the third regime about the mean path when the end points are one standard deviation apart. It should now be intuitively clear why the breakpoints in the third regime of Figure (4.2) were clustered towards the end. The end points of the regimes were drawn from a distribution with a standard deviation exceeding 60 while the maximum value of one standard deviation within the regime was six. To obtain a breakpoint towards the middle of the regime

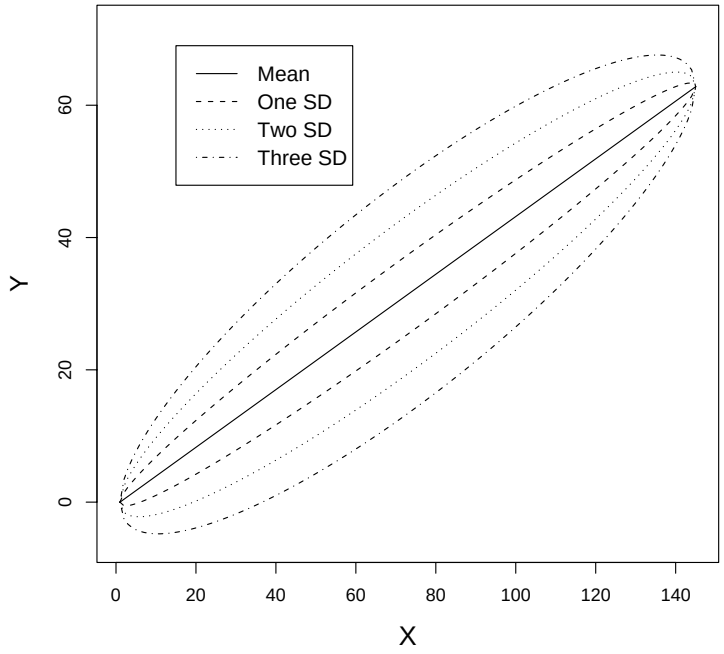


Figure 4.5: Distribution of confidence intervals for the RSS in the third regime with end points one standard deviation apart.

requires that the RSS when the series is split at the start and end of regime three to be very similar in value.

Figure (4.6) presents the predicted breakpoints obtained from 10,000 replications based on the analysis of the third regime in this section.

## 4.4 Reconsidering the Splits

In the literature Quinlan's Error-based Pruning (ERP) (Quinlan, 1993, pp37-41) introduced an innovation in the growing and pruning strategy. ERP prunes back nodes and branches as in other tree selection methods. In addition, it may simplify a tree by grafting a sub-tree $T_t$ into the place of the parent node of $t$. A critique of this method

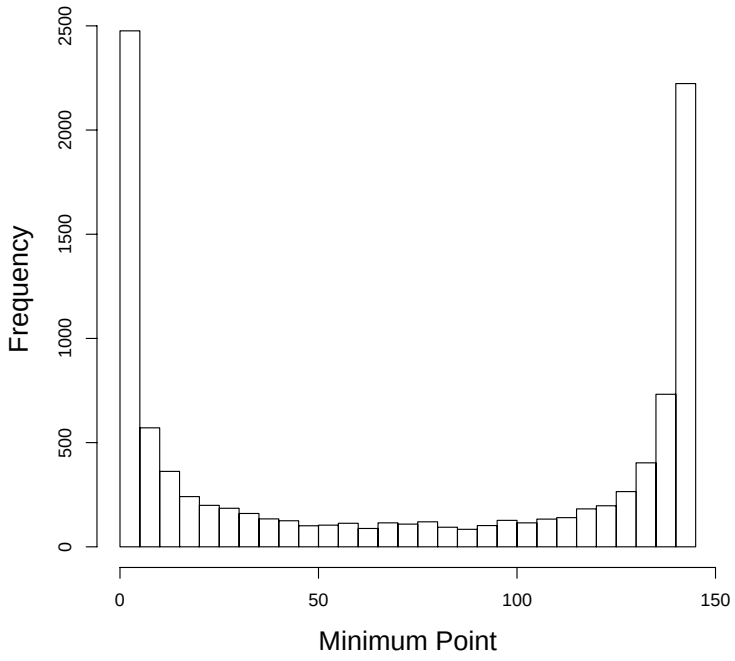


Figure 4.6: Distribution of lowest RSS in the third regime for 10,000 replications using the analysis in Section (4.3).

is contained in Esposito et al. (1997).

We examined this method as a possible way of solving the problem of ART incorrectly splitting the root node (see Section 4.2). It cannot produce the desired improvements because the root node was being split in such a way that two structural breaks were contained in both first level child nodes. Having incorrectly split the root, the tree should contain a minimum of six terminal nodes. Grafting either the left or right subtree back onto the root means half the breaks will be missed.

The problem of selecting a good regression tree has been considered as a Bayesian search by Chipman et al. (1998) and Denison et al. (1998). Their methods overcame some of the limitations of *greedy* algorithms but at the cost of discarding their speed. Chipman et al. (1998) reported their Bayesian search would often get trapped in a

locally good solution and needed to be restarted to get it to consider other, possibly better, trees. While their ideas are interesting, in the context of finding structural breaks in time series they appeared to have no compelling advantages over the BP.

We propose the following algorithm.

1. Grow and prune a regression tree in the usual manner.

2. For each pair of adjacent terminal nodes which do not have the same parent node, use the BIC (or other information criterion if desired) to test whether they should be combined.

3. If so, adjust the split point in the first common ancestor node and combine the nodes.

The algorithm can be applied until no further terminal nodes are combined. We call this algorithm Enhanced Temporal Pruning (ETP) as it may combine nodes adjacent in time which would not be considered for combining in the unmodified scheme.

## 4.5 Improvement

We selected 1,000 trees from each of three different regime lengths which had an incorrect split in the root node but no other incorrect splits and applied the algorithm outlined in the previous section. The results are presented in Table (4.2). This is a worthwhile improvement in the tree pruning algorithm for little additional computational cost.

| Regime Length | Percent Repaired |
|---:|---:|
| 144 | 84.3 |
| 256 | 85.3 |
| 400 | 85.2 |

Table 4.2: Percentage of trees with an incorrectly split root node which were successfully repaired by Enhanced Temporal Pruning in the noisy square wave simulations.

# Chapter 5

# Crest Toothpaste Data

In this chapter we briefly review past literature and then apply ART, CUSUM, and the BP to the Crest toothpaste data set. These data originate from the Market Research Council of America's panel of household purchase records for the period January 1958 to April 1963. The data were gathered weekly giving a total of 275 observations. It has been studied by a number of authors such as Jones (1970), Palda and Blair (1970), Montgomery (1971), Shuchman and Riesz (1975), Wichern and Jones (1977), and Tsay (1987).

Crest was introduced to the American market in 1956 and was the second most popular toothpaste after Colgate-Palmolive's Colgate brand for the next four years. On 1

Table 5.1: Break dates reported by ART, the BP and the 95% confidence interval for the BP for the Crest toothpaste market share data for the weeks between January 1958 and April 1963.

| ART | BP | BP CI |
|---|---|---|
| 70 | 70 | 62-71 |
| 135 | 135 | 134-137 |
| 207 | - | - |

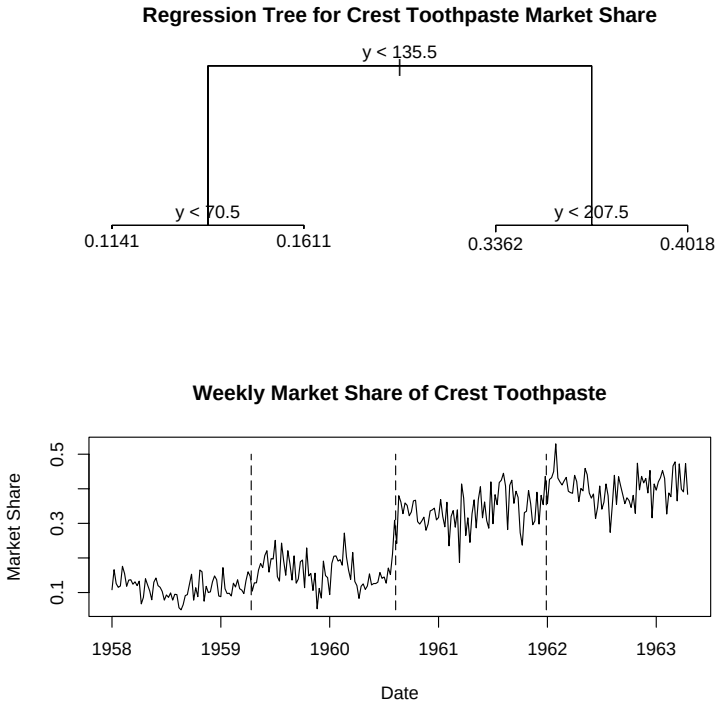


Figure 5.1: The top panel is the regression tree for Crest data showing the breaks at weeks 70, 135, and 207. The bottom panel is the plot of Crest data with the same breaks reported by ART marked by dashed vertical lines.

August 1960 the Council on Dental Therapeutics of the American Dental Association (ADA) made an endorsement of Crest as an aid to dental hygiene after determining that Crest's active ingredient, stannous fluoride, did reduce dental caries. Crest's manufacturer, Proctor and Gamble, "reintroduced" the product and in the next few weeks Crest's market share rose dramatically. Despite vigorous advertising campaigns and special promotions by Colgate and other brands, Crest's market share continued to rise.

Advertising expenditure data are available on a annual basis. Shuchman and Riesz (1975) noted that Crest's market share continued to climb in 1962 even though Crest's advertising expenditure declined from \$10,545,000 to \$9,220,000 between 1961 and 1962

after an increase from $6,242,000 in 1960.

Shuchman and Riesz (1975) state that after the ADA endorsement, the market share of Crest rose to about 35%. After this initial dramatic rise in market share in week 138, Crest's market share did not return to this level again until week 168, more than half a year later. This appears to correspond to the period of intensive counter-advertising studied by Shuchman and Riesz (1975), and "deals" offered by the other manufacturers studied by Montgomery (1971).

Montgomery (1971) and Shuchman and Riesz (1975) studied a number of social, demographic and psychological factors of those who switched to Crest in the post-endorsement period. However, they did not study differences between early and late adopters even though both sets of authors noted the continued rise in Crest's market share after the initial jump following the ADA endorsement.

When we analysed the data with ART, the regression tree indicated that the increase in market share reported by these authors occurred in two steps (see Figure 5.1). The first occurred in early August 1960 at week 135 immediately after the ADA endorsement to a market share of about 34%. The second was an increase to about 40% market share from the beginning of 1962 at week 207. This two-step rise was also seen in the recursive use of the established CUSUM test (see Figure 5.2). This was essentially a manual application of recursive partitioning using the CUSUM test as the splitting criteria.

However, the BP reported only two breaks (see Table 5.1) which exactly matched the first two breaks reported by ART. The failure to report a break at or near the 207th week was consistent with our simulation experiments in which the BP rarely reported too many, but occasionally missed, breaks.

Although all three tests report a break at or near week 70 the literature does not suggest an underlying reason for this. The break at week 135 was clearly attributable to the ADA endorsement. The break at week 207 reported by ART and CUSUM was commented on by various researchers only as a subjective assessment of a visual inspection of the data that Crest's market share was continuing to rise. This shows the

need for structural break methods, such as ART, to indicate when substantial changes have taken place in a time series and to allow a fuller investigation of the underlying reasons to be undertaken.

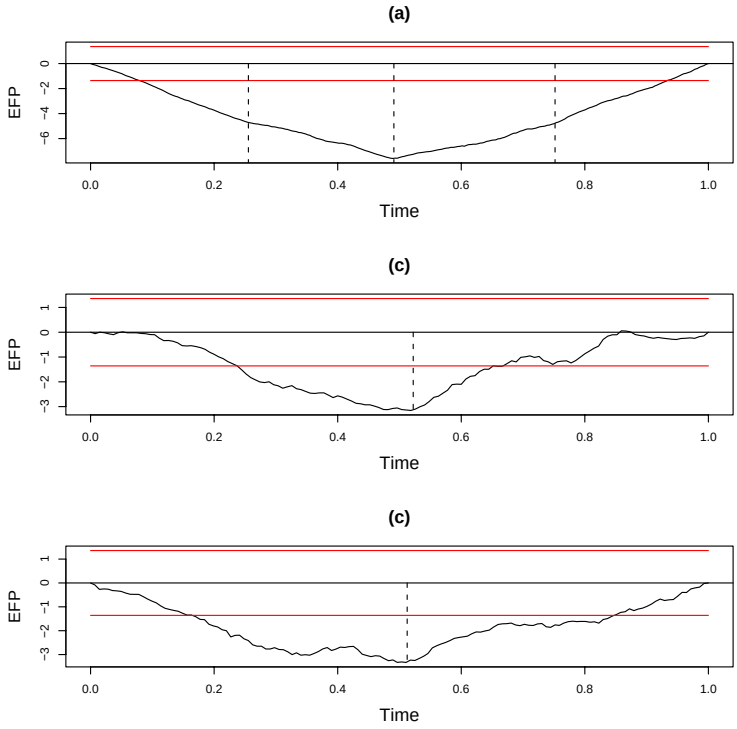


Figure 5.2: Panel (a) is the CUSUM plot for the whole series of 275 observations. Panel (b) is the CUSUM plot from the start of the series to the break in early August 1960 at week 135. Panel (c) is the CUSUM plot from the break in August 1960 to the end of the series. The dashed lines are the ART breakpoints. The thin lines about the heavier zero line are the 5% significance level.

# Chapter 6

# Long Memory Time Series

## 6.1 Introduction

In this chapter we present a summary of long-memory time series which are relevant to the two remaining case studies in this thesis.

Long memory, sometimes called global dependence, long-range dependence, strong dependence or the Hurst phenomena, has been observed in many data sets from such diverse fields as agronomy, climatology, economics, finance, hydrology, and telecommunications. A number of mathematical models have been developed to capture the essential features of this phenomena. Some models can truly be said to have long memory. The most popular true long memory models are the fractional (or fractal) Gaussian noises (FGNs) introduced into applied statistics by Mandelbrot and van Ness (1968), and the fractionally integrated processes of order $d$ (FI(d)) introduced independently by Granger and Joyeux (1980) and Hosking (1981). Both FGNs and FI(d)s are discussed in Section (6.3) below.

FGNs and FI(d)s have been extensively studied. See the volumes by Beran (1994), Embrechts and Maejima (2002), and Palma (2007) and the collections of Doukhan et al. (2003) and Robinson (2003) and the references therein.

Despite a great deal of mathematical work the physical cause or causes of long memory remain unknown for many series.

## 6.2 Properties and Tests for Long Memory

This section and the next draw heavily on the monograph of Beran (1994).

Beran (1994) suggested that data sets which exhibit long memory have three common qualitative features. They are:

1. There are relatively long periods where the observations tend to be either above or below the global mean.

2. When examining short subseries there appear to be local cycles or trends, but no discernible persisting cycle or trend over the whole series.

3. Globally the series look stationary.

It is worth remarking that (1) above could be taken to suggest a structural break model may be appropriate.

If we assume the process is stationary, otherwise the expected value of the mean ($E(\bar{X})$) is not a constant, there are three mathematical properties exhibited by long memory series.

The first involves the variance of the mean of the sampling distribution. For an independent random sample of size $n$ drawn from an infinite population with mean $\mu$ and variance $\sigma^2$ the variance of the sample mean is

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}. \tag{6.1}$$

We can think of Equation (6.1) as a rate of convergence of the variance of the sample mean, $\bar{X}$, to zero, the rate being $1/n$. In time series we are usually dealing with samples for which the correlation is non-zero. This affects the size of the variance of the sample mean. It can be shown (see Chatfield, 2004, p58) that for a sample size $n$ drawn from a population with mean $\mu$, variance $\sigma^2$ and an ACF $\rho(k)$ that

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\left(1 + 2\sum_{k=1}^{n-1}\big(1 - \frac{k}{n}\big)\rho(k)\right). \tag{6.2}$$

For short-memory processes, for all practical purposes, there is some $K$ for which we can regard $\rho(k) = 0; k > K$. More technically, the terms in the brackets in Equation

(6.2) converge to a finite limit, often quite rapidly. Once the terms in the brackets on the right hand side of Equation (6.2) have converged to within some $\epsilon > 0$ of the limit, which in practice is indistinguishable from 0, the convergence proceeds as $1/n$ as in the uncorrelated case Equation (6.1). We can write Equation (6.2) as

$$\mathrm{Var}(\bar{X}) \approx \frac{\sigma^2}{n} c(\rho) \tag{6.3}$$

where $c(\rho)$ is a constant for any given autocorrelation function when the number of observations exceeds $K$.

With long-memory time series the variance of the sample mean appears to differ from Equation (6.1) not only by a constant factor, as we have written in Equation (6.3), but also by a slower rate of convergence to zero than $1/n$. The convergence rate is $1/n^{\alpha}; \alpha \in (0, 1)$, thus Equation (6.3) needs to be generalized to

$$\mathrm{Var}(\bar{X}) \approx \frac{\sigma^2}{n^{\alpha}} c(\rho). \tag{6.4}$$

This leads us to the first indication of the presence of long memory. We can plot the log of the variance of the sample mean against the log of the sample size for a number of different sized subsamples of the series. The slope gives us an estimate of the value of $\alpha$.

The second test involves the autocorrelations $\rho(k)$ (see Beran, 1994, p6). In Equation (6.4) the form of $c(\rho)$ is

$$c(\rho) = \lim_{n \to \infty} \frac{1}{n^{2-\alpha}} \sum_{i \neq j} \rho(i, j)$$

where $\alpha \in (0, 1)$ as before. If we confine ourselves to autocorrelations which only depend on the lag, $|i - j|$, then from Equations (6.3) and (6.4) we can see that for increasing sample size the sum of all correlations from lag $-(n-1)$ to $n-1$ must be proportional to a constant times $n^{1-\alpha}$. That is

$$\sum_{k=-(n-1)}^{n-1} \rho(k) \approx \text{const} \times n^{1-\alpha}. \tag{6.5}$$

As $\alpha \in (0,1)$, this implies

$$\lim_{n\to\infty} \sum_{k=-(n-1)}^{n-1} \rho(k) \to \infty.$$

That is to say, the autocorrelations decay to zero so slowly that their sum diverges to infinity. If Equation (6.5) holds, then asymptotically

$$\lim_{k\to\infty} k^{\alpha}\rho(k) \to c_{\rho}.$$

This is difficult to test for in series of the length typically found in financial or geophysical applications. It is instead a subjective assessment of the ACF for many lags. Fortunately there is a counter-part in the frequency domain which brings us to our third mathematical property.

If we recall that the spectrum of a time series is given by

$$f(\omega) = \frac{\sigma^2}{2\pi} \sum_{k=-\infty}^{\infty} \rho(k) e^{ik\omega}$$

(see Chatfield, 2004, p113) then

$$\lim_{\omega\to 0} f(\omega) \to \infty. \tag{6.6}$$

Thus we can assess whether there is long memory in an apparently stationary finite series by looking at the behaviour of the estimated spectrum at low frequencies. If the spectral estimate tends towards infinity it may signal the presence of long memory.

## 6.3 Models with Long Memory

In this section we outline the basic features of the two most common models with genuine long-memory properties.

### 6.3.1 Fractional Gaussian Noise

**Definition** A real-valued stochastic process $X = \{Z(t)\}_{t\in\mathcal{R}}$ is self-similar with index $H > 0$ ($H$-self-similar) if, for any $a > 0$

$$\{Z(at)\}_{t\in\mathcal{R}} =_d \{a^H Z(t)\}_{t\in\mathcal{R}}$$

where $=_d$ denotes equality in distribution and $\mathcal{R}$ denotes the real numbers.

The index $H$ is called the self-similarity parameter, the scaling exponent or the Hurst parameter in honour of H. E. Hurst who brought the phenomena to prominence and developed the R/S statistic (see Section 7.2 below).

**Definition** A real-valued process $\{Z(t)\}_{t\in\mathcal{R}}$ has stationary increments if, for all $h \in \mathcal{R}$,

$$\{Z(t+h) - Z(h)\}_{t\in\mathcal{R}} =_d \{Z(t) - Z(0)\}_{t\in\mathcal{R}}.$$

H-self-similar processes with stationary increments ($H$-sssi) give rise to stationary sequences with long memory. For a time series $\{X_j\}$, if $\{Z(t)\}_{t\in\mathcal{R}}$ is an $H$-sssi process then

$$X_j = \Delta Z(j) = Z(j+1) - Z(j)\ ,\ j \in \mathcal{Z} \tag{6.7}$$

where $\mathcal{Z}$ denotes the integers, is a stationary sequence.

**Definition** A Gaussian $H$-sssi process $\{B_H(t)\}_{t\in\mathcal{R}}$ with $0 < H < 1$ is called fractional Brownian motion (FBM). It is called standard if $\sigma^2 = \text{var}[B_H(1)] = 1$.

**Definition** If $\{Z(t)\}_{t\in\mathcal{R}}$ is an FBM, then the increment of the process (Equation 6.7) is called fractional Gaussian noise (FGN).

Taqqu (2003) states that the increment sequence $\{X_k\}_{k\in\mathcal{Z}}$ has the following properties:

1. $\{X_k\}_{k\in\mathcal{Z}}$ is stationary.
2. $E[X_k] = 0$.
3. $E[X_k^2] = \sigma^2 = E[Z(1)^2]$.
4. The autocovariance function of the process $\{X_k\}_{k\in\mathcal{Z}}$ is given by

$$\begin{aligned}\gamma(k) &= E[X_i X_{i+k}] \\ &= \frac{\sigma^2}{2}\left(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}\right) \\ &= \frac{\sigma^2}{2}\Delta^2|k|^{2H}\end{aligned}$$

where $\Delta^2$ denotes the second difference.

5. For $k \neq 0$

$$\gamma(k) = \begin{cases} < 0, 0 < H < \frac{1}{2} \\ = 0, H = \frac{1}{2} \\ > 0, \frac{1}{2} < H < 1. \end{cases}$$

6. If $H \neq 1/2$, then

$$\gamma(k) \approx \sigma^2 H(2H-1)|k|^{2H-2} \text{ as } k \to \infty \tag{6.8}$$

From Equation (6.8) $\gamma(k)$ tends to zero like a power function with increasing $k$. When $1/2 < H < 1$ it tends to zero so slowly that the sum

$$\sum_{k=-\infty}^{\infty} \gamma(k)$$

diverges. Thus the process represented by Equation (6.7) displays long-memory.

It can be shown (Beran, 1994, p53) that the spectral density of FGN is

$$f(\omega) \sim c\omega^{1-2H}$$

where $\omega$ is the frequency and $c$ is a constant. Thus the long-memory property corresponds to a divergence of the spectral representation to infinity at the origin. Of course, in finite samples $f(0)$ is finite.

Some authors were enthusiastic about the ability of FGNs to model hydrological time series. Wallis and O'Connell (1973) wrote "It has been exhaustively documented that discrete time approximations to fractional Gaussian noise provides a necessary and sufficient explanation of the Hurst phenomenon". On the other hand, Hipel and McLeod (1978) reported that FGNs were inferior to several other processes in modeling long memory series, in particular to the Box-Jenkins ARMA models.

### 6.3.2 ARFIMA Models

The other common true long memory model is the extension of Box-Jenkins ARIMA(p,d,q) models to non-integer values of $d$ which was accomplished independently by Granger

and Joyeux (1980) and Hosking (1981). Several factors motivated these authors to consider fractional differencing.

For a stationary short memory time series the sample ACF decays exponentially as the lag increases. For a non-stationary random walk or unit-root process the sample ACF converges to one for all fixed lags as the sample size increases. For series with a unit-root, taking first differences yields a stationary time series with an ACF which decays exponentially as before. Intermediate between these are long memory time series which have a sample ACF which decays at a polynomial rate with increasing lag and for which taking first differences yields a series which appears to be over-differenced. Thus some model with an order of differencing intermediate between zero and one seemed required.

Hosking (1981) attributed the above observation of Hipel and McLeod (1978) to FGNs' inability to model low-lag correlation structures correctly and stated this motivated his research when he added non-zero AR and MA orders to his fractional differencing model.

These ARIMA(p,d,q) models with non-integer $d$ values are usually called AutoRegressive Fractionally Integrated Moving Average (ARFIMA) series and are normally defined for $-1/2 < d < 1/2$ as any other series can be differenced until $d$ lies in this range. When $-1/2 < d < 0$ the series is anti-persistent. Henry and Zaffaroni (2003) state that anti-persistent series are characterized by a shrinking spectral density towards zero frequency.

Recall (Chatfield, 2004, p48) that an ARIMA(p,d,q) model is defined as

$$\phi(B)(1-B)^d X_t = \theta(B)\epsilon_t \tag{6.9}$$

where $B$ is the backward shift operator ($B[X_t] = X_{t-1}$), and $\phi(B)$ and $\theta(B)$ are polynomials in $B$ of order $p$ and $q$ respectively, $X_t$ is the observation at time $t$ and the $\epsilon_t$ are usually assumed to be a white noise sequence drawn from an $N(0,1)$ distribution.

If, for simplicity, we assume $\phi(B) = \theta(B) = 1$ then Equation (6.9) reduces to

$$(1-B)^d X_t = \epsilon_t \tag{6.10}$$

or

$$X_t = (1 - B)^{-d}\epsilon_t. \tag{6.11}$$

If $-1/2 < d < 1/2$ then Equation (6.10) can be written as an infinite order AR process (Beran, 1994, pp64,65)

$$\sum_{k=0}^{\infty} \pi_k X_{t-k} = \epsilon_t \tag{6.12}$$

where

$$\pi_k = \frac{\Gamma(k-d)}{\Gamma(k+1)\Gamma(-d)}$$

and $\Gamma(\cdot)$ is the gamma function.

Similarly, Equation (6.11) can be written as an infinite order MA process

$$X_t = \sum_{k=0}^{\infty} a(k)\epsilon_{t-k} \tag{6.13}$$

where

$$a(k) = \frac{\Gamma(k+d)}{\Gamma(k+1)\Gamma(d)}.$$

It can be shown (Beran, 1994, pp63-64) that the covariance function is given by

$$\gamma(k) = \sigma_\epsilon^2 \frac{(-1)^k \Gamma(1-2d)}{\Gamma(k-d+1)\Gamma(1-k-d)}$$

and the correlations are equal to

$$\rho(k) = \frac{\Gamma(1-d)\Gamma(k+d)}{\Gamma(d)\Gamma(k+1-d)},$$

and

$$\rho(k) \sim \frac{\Gamma(1-d)}{\Gamma(d)}|k|^{2d-1} \text{ for } |k| \to \infty.$$

Thus for $0 < d < 1/2$ a fractionally integrated process exhibits long memory as all previous states of the process influence the present.

In the last paragraph of Hosking (1981) he mentions in passing that fractional differencing may prove useful with the process $(1 - 2\phi B + B^2)^d x_t = \epsilon_t$, with $|d| < 0.5$ and $|\phi| < 1$. This bears a striking resemblance to the so-called Gegenbauer process which is elaborated on in Beran (1994, pp213–215). Gegenbauer processes exhibit both long memory and behaviour which is approximately cyclic but not strictly periodic.

## 6.4 Short Memory Models

FGNs and FI(d) series can be considered to be true long memory models for reasons outlined above. Other models display the long memory property but are, in fact, short memory series. We consider two classes of such models.

### 6.4.1 Aggregation Models

In economic and financial applications many time series are aggregates of numerous individual component series. For example, stock market indices, such as the S&P 500, are just a (possibly weighted) sum of a few tens to thousands of individual stocks. Granger and Morris (1976) examined the types of series which resulted from summing a number of AR(p) and MA(q) series. Granger and Morris (1976) considered the AR and MA models to be "...intuitively reasonable and could well occur in practice ...". Of particular interest, they reported that the sum of an AR($p_1$) and an AR($p_2$) series yielded an ARMA($p_1 + p_2, \max(p_1, p_2)$) series.

Subsequently Granger (1980) considered whether the sum

$$X_t = \sum_{j=1}^{N} X_t(j)$$

where $X_t(j); j = 1, 2, \ldots, N$ are the individual components of an aggregated series $X_t$, each of which has short-memory, could exhibit long memory. Granger reported this was the case. Beran (1994, p16) stated "...observing long-range dependence in an aggregated series (macro-level) one cannot conclude that this long memory is due to the genuine occurrence of long memory in the individual series (micro-level). It is possible that instead it is induced artificially by aggregation."

Aggregation models have also appeared in the non-economic literature, albeit indirectly. Hare and Mantua (2000) amassed 100 North Pacific climatic and biological time series and used a composite analysis to demonstrate regime shifts in the Pacific Decadal Oscillation (PDO). Rudnick and Davis (2003) responded with a paper criticizing their methodology and demonstrating that composite analysis of red noise series produced

the step-like changes seen in the PDO. Overland et al. (2006) produced a third paper which suggested a long-memory model also fitted the data. It would appear all three sets of authors were discussing, in part, a Granger-style long memory aggregation model without being aware of his work.

Cioczek-Georges and Mandelbrot (1995) have studied an alternative aggregation model which exhibits long memory properties. Their model has found application in modeling aggregate traffic on computer networks. Willinger et al. (2003) also studied long memory in computer networks from the point of view of the network protocols which govern how traffic from individual sources is aggregated onto shared data networks.

### 6.4.2 Constrained Non-Stationary Models

Klemes (1974) argued that long memory in hydrological time series was a statistical artifact caused by analyzing non-stationary time series with statistical tools which assume stationarity. Often series which display the long memory property are constrained for physical reasons to lie in a bounded range. Beyond that we have no reason to believe that they are stationary. For example, tree ring indices often display long memory (see Chapter 10). For biological reasons there are fixed upper and lower bounds on the rate of tree growth. Long periods of above or below average growth may have biological reasons ultimately rooted in climatic shifts and may not be meaningless fluctuations in growing conditions.

To see how confusion between a non-stationary series and true long memory may arise consider the two ACFs presented in Figure (6.1). The ACF in panel (a) shows no statistically significant autocorrelations as it is the ACF of a series of random numbers drawn from an N(0,1) distribution. The ACF in panel (b) appears to be of the long memory type yet it is the ACF of the same series with a single mean shift at the mid-point.

It is mathematically difficult to handle a general constrained non-stationary model. To make the problem tractable the models which have been studied have often been

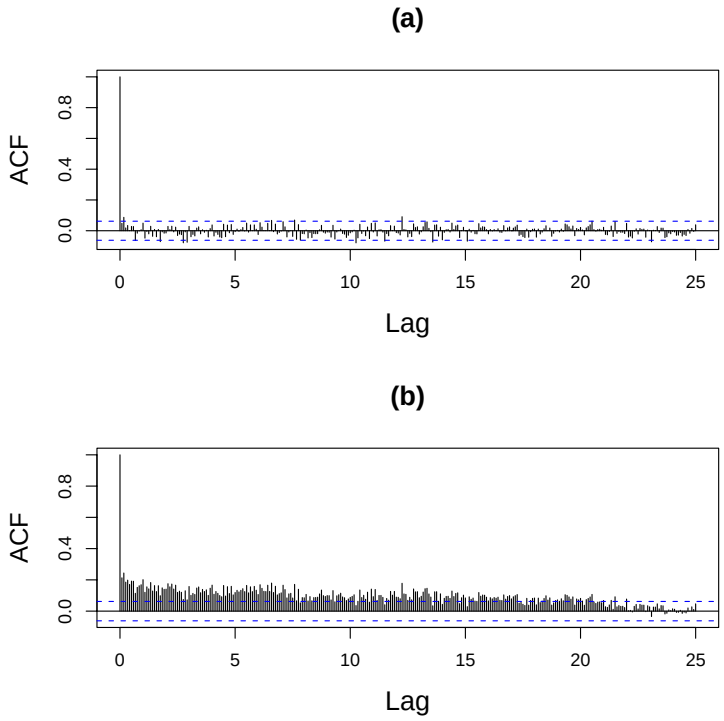


Figure 6.1: The ACF of two simulated time series. Panel (a) is the ACF of random numbers with no serial correlation. Panel (b) is the ACF of the same data with a mean shift at the mid-point.

of the type where the system has two or more well defined states which it switches between on some irregular basis. Often the models studied are stationary but are used as simplified models of the more general non-stationary series. Thus a number of authors have studied regime switching and hidden Markov models. One of the simplest is due to Granger and Teräsvirta (1993) who introduced the regime switching model

$$x_t = \text{sgn}(x_{t-1}) + \epsilon_t$$

where $\epsilon_t$ is *iid* $N(0, \sigma_\epsilon^2)$ and

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases}$$

The process $x_t$ is Markovian, with zero mean and is stationary. They observed empirically that for infrequent regime changes the ratio $\log(|\rho_k|)/\log k$ was constant with increasing k. This was the same theoretical form as for an FI(d) series. In this case the model exhibited statistical long memory in the time domain even though it was Markovian.

Granger and Hyung (2004) showed that the model

$$x_t = \mu_t + \epsilon_t \tag{6.14}$$

$$\mu_t = \mu_{t-1} + q_t \eta_t \tag{6.15}$$

where

$$t = 1, \ldots T$$

$$\mu_t = \text{mean level of the series at time t}$$

$$\epsilon_t = \text{a noise variable}$$

$$\eta_t = N(0, \sigma^2)$$

$$q_t = \begin{cases} 0, \text{with probability } 1 - p \\ 1, \text{with probability } p \end{cases}$$

and the assumption that the probability of breaks, $p$, converged slowly to zero as the sample size increased, i.e.

$$\lim_{T \to \infty} p \to 0$$

and

$$\lim_{T \to \infty} Tp = \text{non-zero constant}$$

yielded an ACF of the long-memory type.

The purpose of the assumption was to ensure that the expected number of breaks, $Tp$, was bounded from above as $T$ increased to infinity. To preserve the long memory property the number of breaks was required to remain finite as the sample size increased.

The assumption made by Granger and Hyung (2004) was the same as that used to obtain the Poisson distribution from the binomial (Miller and Miller, 1999, pp186,187). Further, the model (6.14) and (6.15) is a small variation on the random walk plus noise model (Chan, 2002, p139) often used in the discussion of the Kalman filter. The difference was that the assumption of Granger and Hyung (2004) limited the random walk to a finite number of steps.

A more general model is the structural break model which was defined in Equations (1.1) and (1.2) in Section (1.1).

## 6.5 The Beran Test

FGNs and FI(d) series with long-range dependence have a distinctive spectral density at the origin of the form

$$f(\omega) \sim b|\omega|^{1-2H} \tag{6.16}$$

where $b$ is a constant and $H \in (1/2, 1)$. By building on the work of Milhoj (1981), Beran (1992) extended the goodness-of-fit test in the frequency domain to Gaussian models with spectra of the form in Equation (6.16).

The null hypothesis was that the data had a spectral density of the form of Equation (6.16). There are a number of reasons why a process may appear to have long memory but whose spectral density differs from Equation (6.16). Thus the alternative was simply that the process had some spectral density other than Equation (6.16). Consequently a rejection of the null does not lead to a clear alternative.

Beran stated that the test was especially powerful against a model misspecification which underestimated the parameter $H$ in Equation (6.16).

The test statistic estimated the departure from the modeled spectral density over

the whole range $[-\pi, \pi]$ and is given by

$$T_n(\theta) = \frac{4\pi}{n} \sum_{j=1}^{n^*} \left( \frac{I(\omega_j)}{f(\omega_j)} \right)^2 / \left( \frac{4\pi}{n} \sum_{j=1}^{n^*} \frac{I(\omega_j)}{f(\omega_j)} \right)^2$$

where

$$\begin{aligned}
\theta &= \text{vector of parameters for which the goodness of fit is sought} \\
\omega_j &= \frac{2\pi j}{n} \quad (j = 1, 2, \ldots, n^*) \text{ are the Fourier frequencies} \\
n^* &= \begin{cases} (n-1)/2 \text{ if } n-1 \text{ is even} \\ (n-1)/2 - 1/2 \text{ if } n-1 \text{ is odd} \end{cases} \\
I &= \text{the series periodogram} \\
f &= \text{the spectral density of the model being tested against.}
\end{aligned}$$

Alternatively the test can be written as

$$(2\pi)^{-1} \sum_{k=0}^{n-1} (\hat{\gamma}_k / \hat{\gamma}_0)^2$$

where the $\hat{\gamma}_k$ are the estimated covariances of the residual process. This seems particularly appropriate for long memory processes with their very slow decay of the serial correlations.

In the remainder of the thesis we applied the Beran test using functions implemented in the `R` package `longmemo` of Beran et al. (2006). Clear instructions on its use were provided in the manual. It is simple to perform and among the results reported is a $p$-value indicating the goodness-of-fit.

# Chapter 7

# $H$-Estimator Selection

## 7.1 Introduction

In this chapter we review some of the literature on the estimators of $H$ and $d$, perform an empirical study on simulated series with lengths in the range of the data sets we subsequently examine in the case studies, and select estimators for routine use in the remainder of this thesis. This chapter became necessary because the various estimators of $H$ and $d$ were behaving inconsistently when applied to both real and simulated data. Apart from the papers of Taqqu et al. (1995) and Jeong et al. (2007) there appeared to be no other comparative empirical study of $H$ and $d$ estimators, although there had been considerable theoretical work done on them.

In the analysis of long-memory time series it is of critical importance to estimate the strength of the long-range dependence. The two common measures $H$ and $d$ were defined and discussed in Chapter (6).

A number of estimators of $H$ and $d$ have been developed. These are usually validated by an appeal to some aspect of self-similarity, or by an asymptotic analysis as the length of the time series diverges to infinity. Taqqu et al. (1995) carried out an empirical study of nine estimators for a single series length of 10000 observations, five values of both $H$ and $d$, and 50 replications. Since then computer power has increased considerably. Jeong et al. (2007) carried out a comparison of six estimators on simulated FGNs with

32768 ($2^{15}$) observations, four values of $H$ and 100 replications. We have extended these studies to a larger range of parameters, two additional estimators, and a higher number of replications as detailed in Section (7.3) below.

## 7.2 The Estimators

In this section we anticipate the results of the empirical study presented in Section (7.4) below and only give a detailed description of the Whittle estimator. For the remainder we give a brief description and point to the relevant literature where they have been studied. A number of these estimators are treated in detail in the volume of Beran (1994) chapters 4 through 6.

**Variance Methods** The aggregated variance method, also known as the variance plot, is the basis for three estimators, if the Higuchi (1988) estimator is excluded. The aggregated variance method utilizes the fact that for long-memory processes the variance of the sample mean converges to zero slower than $1/n$, thus exploiting the property given in Equation (6.4).

For a long-memory process Beran (1989) proved

$$\mathrm{Var}(\bar{X}_n) \approx Cn^{2H-2}$$

where $C > 0$ is some constant. This suggests $H$ can estimated by examining the rate of convergence towards zero of the variance of the sample mean of the series. The procedure is as follows.

For different integers in the range $2 \leq k \leq n/2$ and a sufficient number, $m$, of subseries of length $k$, calculate the sample means

$$\bar{X}_1(k), \bar{X}_2(k), \ldots, \bar{X}_{m_k}(k)$$

and the overall mean

$$\bar{X}(k) = \frac{1}{X_{m_k}} \sum_{k=1}^{m_k} \bar{X}_j(k).$$

For each value of $k$, calculate the variance of the sample means

$$s^2(k) = \frac{1}{m_k - 1} \sum_{k=1}^{m_k} (\bar{X}_j(k) - \bar{X}(k))^2.$$

Then plot $\log s^2(k)$ against $\log k$. The slope of the line can be estimated by least-squares regression. For an FGN the slope is $2H - 2$.

In the absolute value method instead of computing the sample variance, the sum of the absolute values of the aggregated series is calculated.

$$\frac{1}{\lfloor N/m \rfloor} \sum_{k=1}^{\lfloor N/m \rfloor} |x^{(m)}(k)|$$

where

$$x^{(m)}(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X(i); k = 1, 2, \ldots$$

If the series is an FGN then the following scaling law applies

$$\frac{1}{\lfloor N/m \rfloor} \sum_{k=1}^{\lfloor N/m \rfloor} |x^{(m)}(k)| \sim m^{H-1}.$$

The theoretical properties of the aggregated variance and absolute value methods have been investigated by Giraitis et al. (1999). Also see Taqqu et al. (1995).

The differenced variance method is another modification of the aggregated variance method which was developed by Teverovsky and Taqqu (1999) to be robust to trends and shifting means. Deterministic trends are known to cause other estimators to spuriously report $H$ estimates which indicate long memory when, in fact, none exists. See, for example, Bhattacharya et al. (1983) and Smith (2005). The estimator uses the blocks of the aggregated variance method and takes the differences between the blocks. Teverovsky and Taqqu (1999) showed that this removed any effects of shifting means or trends.

**Haslett-Raftery** This method was proposed in the context of spatial and temporal multivariate time series analysis by Haslett and Raftery (1989) and utilized earlier work by Hosking (1981). It is an approximate maximum likelihood estimator of $d$ in the time domain.

**Higuchi** This method was introduced by Higuchi (1988). It is similar to aggregated variance methods. It calculates the length of a path from the time series data and estimates its fractal dimension $D$. If the curve is an FGN then there is a simple relationship between $H$ and $D$, $H = 2 - D$. Wuertz (2005) implemented the Higuchi method to report an estimate of $H$. If a researcher which to obtain an estimate of the fractal dimension it can be easily recovered from the estimated $H$ value.

**Peng** This method was introduced by Peng et al. (1994). It involves breaking the series into blocks of size $m$. Within each block the partial sums

$$Y(i) = \sum_{j=1}^{i} x_j;\ i = 1, 2, \ldots, m$$

are calculated. A least squares regression line is fitted to the $Y(i)$, the residuals are extracted, and their variance calculated. The procedure is repeated for each of the blocks and the sample variances are averaged. Finally, the log of the sample variances of the residuals are plotted against the log of $m$. For an FGN we obtain a line with slope $2H$.

Taqqu et al. (1995) gave a detailed proof that the method of Peng et al. (1994) was an asymptotically unbiased estimator of $H$.

**Periodogram Estimators** These estimators utilize the property that for an FGN or FI(d) process the spectral density has a pole at zero frequency as in Equation (6.6). In short, the spectral density is proportional to $|\omega|^{1-2H}$ as $|\omega| \to 0$. Thus

$$\log f(\omega) \propto (1 - 2H) \log \omega.$$

Thus plotting the log of the periodogram against the log of the frequency should yield a straight line of slope $1-2H$. Then $H$ can be estimated from the estimated slope, $\hat{\beta}$, obtained from a linear regression

$$\hat{H} = \frac{1 - \hat{\beta}}{2}.$$

The boxed periodogram method was developed specifically to deal with the problem of having most of the points used to estimate $H$ on the right-hand side of the log-periodogram plot. This was believed to, possibly, cause bias in the periodogram estimator for reasons outlined in Beran (1994, p133) and Taqqu et al. (1995). To avoid the bias the boxed periodogram divides the frequency axis into a series of logarithmically equally spaced boxes with the exception of a few values on the left. The values in the boxes are averaged and a regression line is fitted to these averaged values.

Periodogram type estimators have been considered in detail by Robinson (1994, 1995) and Lobato and Robinson (1996).

**Rescaled Range** This estimator was originally proposed by Hurst (1951) who sought to find an ideal capacity for a reservoir in the sense that the outflow is uniform, at time $t+h$ the reservoir is as full as it was at time $t$ and it never overflows. Let $X_i$ be the inflow at time $i$, $Y_i = \sum_{i=1}^{j} X_i$ the cumulative flow up to time $j$. Then the ideal capacity can be shown to be

$$R(t,k) = \max_{0\le i\le k}\big[Y_{t+i} - Y_t - \frac{i}{k}(Y_{t+k} - Y_t)\big] - \min_{0\le i\le k}\big[Y_{t+i} - Y_t - \frac{i}{k}(Y_{t+k} - Y_t)\big].$$

$R(t,k)$ is called the rescaled range. To study the properties which are independent of the scale, $R(t,k)$ is standardized by

$$S(t,k) = \sqrt{\frac{1}{k}\sum_{i=t+1}^{t+k}\left(X_i - \bar{X}_{t,k}\right)^2}$$

where

$$\bar{X}_{t,k} = \frac{1}{k}\sum_{i=t+1}^{t+k} X_i.$$

The ratio

$$\frac{R}{S} = \frac{R(t,k)}{S(t,k)}$$

is called the rescaled adjusted range or R/S statistic. For series with long memory plotting the log of the R/S statistic against the log of $k$ yields a line with a slope

which differs from the value 1/2 found in short memory series. Thus $H$ can be estimated using linear regression on these log-log plots.

Theoretical properties of the R/S estimator have been examined by Mandelbrot (1975), Mandelbrot and Taqqu (1979) and Bhattacharya et al. (1983).

**Wavelet** A estimator based on wavelets is somewhat mathematically complicated. For reasons of space we omit a summary. A detailed description can be found in Abry and Veitch (1998) and Jensen (1999).

The application of wavelets to estimating $H$ have also been studied by Abry et al. (1998) and Veitch and Abry (1999). The wavelet estimator is known to be asymptotically unbiased.

### 7.2.1 The Whittle Estimator

Fox and Taqqu (1986) adapted the approach of Whittle (1953) for estimating the parameters of a time series to long memory series, particularly for estimating the Hurst parameter. They showed the Whittle estimator was consistent and asymptotically normal. Dalhaus (1989) established the efficiency of the Whittle estimator for long-range dependent Gaussian series. Most of the following discussion is drawn from Beran (1994).

The basis of the Whittle estimator is to fit a theoretical spectrum $f(\omega;\theta)$ to the observed periodogram

$$I(\omega) = \frac{1}{2\pi n}\left|\sum_{j=1}^{n} x_j e^{ij\omega}\right|^2$$

where $n$ is the number of data points, and $x_j$ is data point $j$, $\omega$ is the frequency, and $\theta$ is a vector of parameters to be estimated.

The Whittle likelihood is given by

$$L_W(\theta) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log f(\omega;\theta)d\omega + \frac{x^t A(\theta)x}{n} \tag{7.1}$$

where $x$ is a vector of data, $x^t$ is its transpose, $A(\theta)$ is an $n \times n$ matrix

$$A(\theta) = \left[\alpha(j-l)\right]_{j,l=1,\ldots,n}$$

and the elements of this matrix are given by

$$\alpha(j-l) = \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\frac{1}{f(\omega;\theta)}e^{i(j-l)\omega}d\omega.$$

We can approximate this integral by the Riemann sum

$$\hat{\alpha}(k) = 2\left[\frac{1}{(2\pi)^2}\sum_{j=1}^{n}\frac{1}{f(\omega;\theta)}e^{ik\omega_{j,n}}\frac{2\pi}{n}\right]$$

where

$$\omega_{j,n} = \frac{2\pi j}{n}; \qquad j = 1, 2, \ldots, n^* \tag{7.2}$$

and $n^*$ is the integer part of $(n-1)/2$. The $\omega_{j,n}$ are the Fourier frequencies for the series.

This allows us to obtain a discrete version of Equation (7.1). Note that it can be written in terms of the periodogram

$$L_W(\theta) = \frac{1}{2\pi}\left[\int_{-\pi}^{\pi}\log f(\omega;\theta)d\omega + \int_{-\pi}^{\pi}\frac{I(\omega)}{f(\omega;\theta)}d\omega\right].$$

From Equation (7.2)

$$\tilde{L}_W(\theta) = \frac{1}{\pi}\left[\sum_{j=1}^{n^*}\log f(\omega_{j,n};\theta)\frac{2\pi}{n} + \sum_{j=1}^{n^*}\frac{I(\omega_{j,n})}{f(\omega_{j,n};\theta)}\frac{2\pi}{n}\right]. \tag{7.3}$$

Minimizing Equation (7.3) is equivalent to minimizing the ratio of $I(\omega_{j,n})$ to $f(\omega_{j,n};\theta)$. Thus

$$\tilde{Q}(\theta) = \sum_{j=1}^{n^*}\frac{I(\omega_{j,n})}{f(\omega_{j,n};\theta)}$$

can be used to estimate the Hurst parameter by minimizing with respect to $\theta$ the vector of parameters we are estimating. For an FGN $\theta$ may simply be $H$ but for an ARFIMA model $\theta$ may include AR and MA components in addition to $d$. It is possible to use some value $m < n^*$. This has the effect of cutting off some of the high frequencies in the data in the estimation.

The Whittle estimator is an approximate MLE method but has the same asymptotic distribution as the exact MLE, see Beran (1994, p112). Of the ten estimators available in the contributed package `fSeries` of Wuertz (2005) only the Whittle estimator is non-graphical.

## 7.3 Method

Of the 11 estimators described in Section (7.2) the Haslett-Raftery is implemented in the contributed package `fracdiff` of Fraley et al. (2006) and the remaining ten implemented in `fSeries` of Wuertz (2005), for the popular statistical software `R`. Taqqu et al. (1995) simulated both FGNs and the corresponding discrete time FI(d) series and reported that each estimator performed similarly whether estimating $H$ in simulated FGNs or $d$ in simulated FI(d)s. For example, if an estimator was biased when estimating $H$ it was also biased in a very similar manner when estimating $d$. Thus, with the exception of the Haslett-Raftery estimator, we only investigated each estimator's performance in estimating $H$ for simulated FGNs. FGNs were generated using the function `fgnSim` in `fSeries`. We ran 1000 replications of simulated FGNs with 100 different lengths and eight different $H$ values. The lengths were between 100 and 10,000 data points in steps of 100. The $H$ values were between 0.55 and 0.90 in steps of 0.05. For each series $H$ was estimated by each of these ten estimators. For each $H$ value and series length we estimated the median, 75% and 95% confidence intervals empirically from the simulated data. The $H$ or $d$ estimates were sorted into ascending order and the median obtained by averaging the 500th and 501st values. Similar calculations were done for the upper and lower values of the 75% and 95% confidence intervals.

For the Haslett-Raftery estimator of $d$ we generated FI(d) series with the function `farimaSim` in `fSeries` over the range 0.05 to 0.40 in steps of 0.05. The other details are the same as above. In the presentation of the results we converted the $d$ estimates to $H$ equivalents to facilitate comparisons among the estimators.

It should be noted that the series used are pure FGNs or FI(d)s and represent ideal conditions for these estimators. In real series with various departures from an ideal FGN the estimators may well have higher variance or different biases than presented here. In other words the results presented here represent a best case scenario, in practice they may well be worse.

## 7.4 Results

To present the results in tabular form would require a very large amount of space. Thus we present them in graphical form. Figures (7.1) through (7.7) present some of the results. Figures (7.1) through (7.6) are presented with a vertical axis covering 1.2 $H$ units to facilitate comparisons among the estimators' standard deviation of their estimates. It should be noted that stationary long memory occurs in the range $0.5 < H < 1.0$. Baillie (1996) states that for $1.0 \leq H < 1.5$ the series are non-stationary but mean reverting while for $0 \leq H \leq 0.5$ the series are anti-persistent. Figure (7.7) presents the mean squared error (MSE) as a function of series length. We report MSE for series lengths greater than or equal to 500 data points. Again the vertical axes all have the same range to facilitate comparisons.

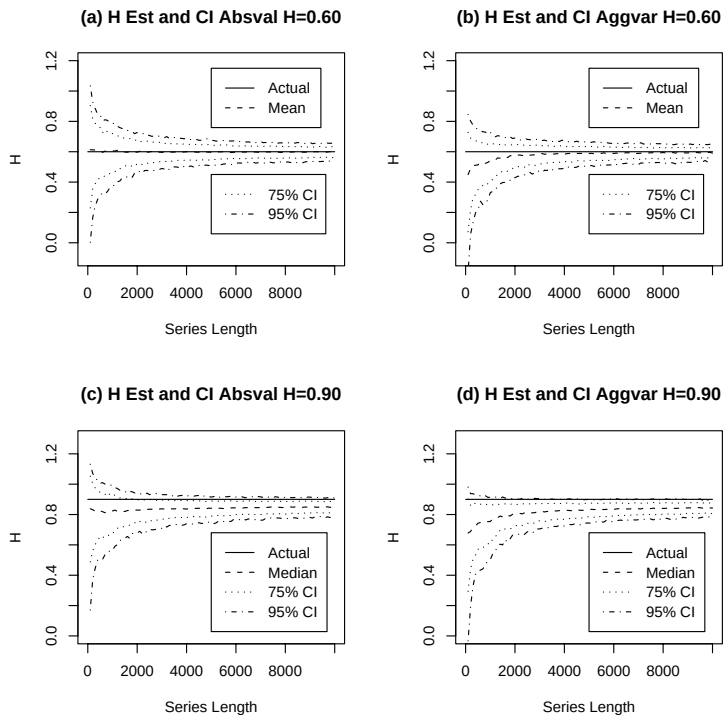


Figure 7.1: Empirical confidence intervals for the $H$ estimates with $H = 0.60$ and $H = 0.90$; (a) and (c) absolute value method, (b) and (d) aggregated variance estimator.

The results for the absolute value of the variance method are presented in Figures (7.1) (a) and (c). The absolute value of the variance method was unbiased at all series lengths when $H$ was low (0.55 or 0.60) but became progressively biased and underestimated $H$ as $H$ increased.

The results for the aggregated variance method are presented in Figures (7.1) (b) and (d). The aggregated variance method exhibited bias and underestimated $H$ in short series when $H$ was low. As $H$ increased the estimator became increasingly biased at all series lengths examined. With $H = 0.90$ the true value of $H$ lay above the upper 95% empirical confidence interval for all but the shortest series lengths.

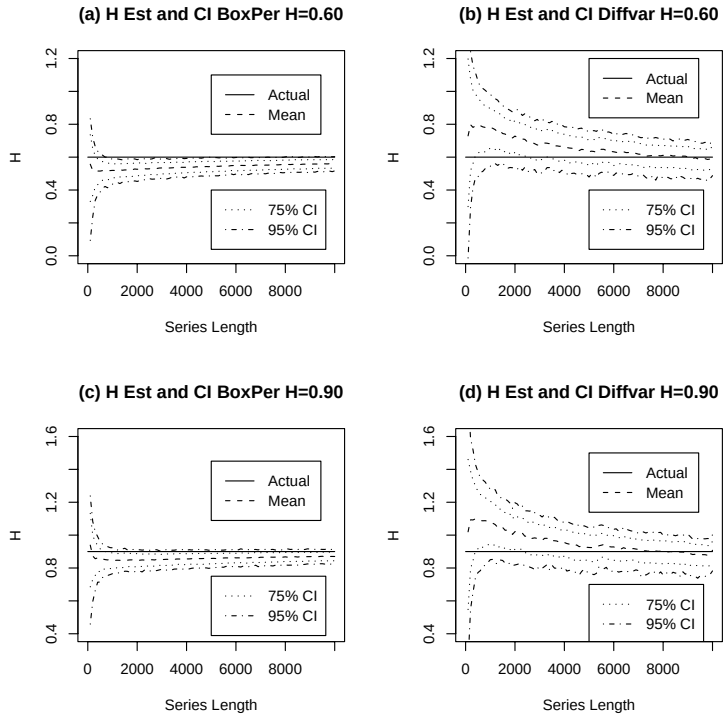


Figure 7.2: Empirical confidence intervals for the $H$ estimates with $H = 0.60$ and $H = 0.90$; (a) and (c) boxed periodogram method, (b) and (d) differenced variance estimator.

The results for the boxed periodogram method are presented in Figures (7.2) (a) and

(c). The boxed periodogram method was developed specifically to deal with perceived problems with the periodogram estimator. Comparing the boxed periodogram with the unmodified periodogram method in Figures (7.4) (a) and (c) we can see that for FGNs where the series were short and $H$ was high, that the periodogram method was biased towards over estimating $H$. The boxed periodogram was biased towards underestimating $H$ for almost all values of $H$ and series lengths examined.

The results for the differenced variance method are presented in Figures (7.2) (b) and (d). The differenced variance method had one of largest confidence intervals of the estimators when the series were short but this slowly decreased as sample size increased. Only the periodogram and wavelet methods had a similarly wide confidence interval for short series. The differenced variance estimator exhibited bias towards over estimating $H$ for any series with less than 7,000 observations. The bias was very serious in the short series. For series longer than about 9,000 observations the estimator exhibited a small amount of bias towards underestimating $H$.

The results for the Higuchi (1988) estimator are presented in Figures (7.3) (a) and (c). The Higuchi was biased towards underestimating $H$ but the magnitude of the bias appeared relatively independent of $H$. The width of the confidence interval of the estimate increased with increasing $H$.

The results for the Peng et al. (1994) estimator are presented in Figures (7.3) (b) and (d). The Peng estimator was biased toward under estimating $H$ in the series lengths we investigated. This bias appeared to be independent of $H$ but greater in short series.

The results for the R/S estimator are presented in Figures (7.4) (b) and (d). The R/S estimator is of considerable historical interest because it was first proposed by Hurst and was used extensively in early studies of long-memory processes. However, as can be seen from Figures (7.4) (b) and (d) the R/S estimator exhibited three problems; it was biased upwards when $H$ was low, it was biased downwards when $H$ was high, and the confidence interval of the estimate did not decrease with increasing series length once the series reached about 1000 observations.

The results for the Whittle estimator are presented in Figures (7.5) (a) and (c).

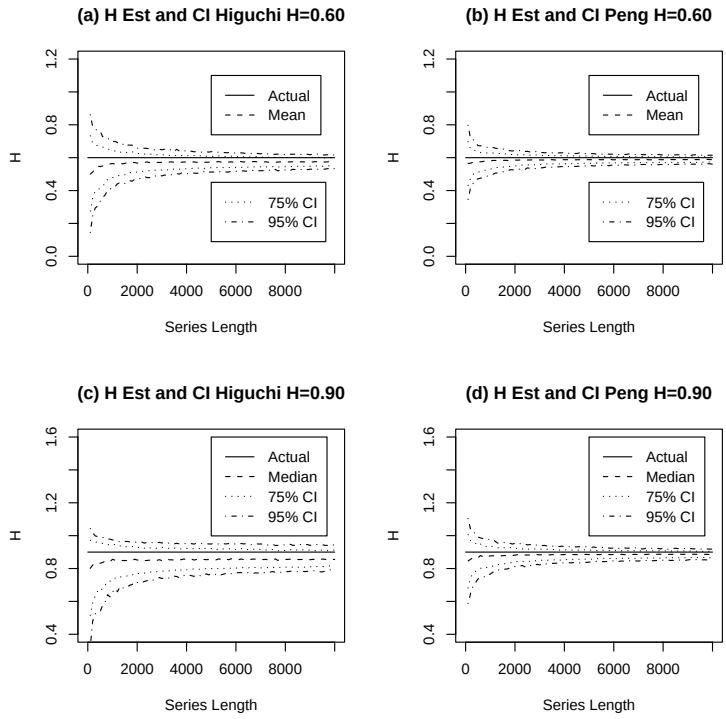


Figure 7.3: Empirical confidence intervals for the $H$ estimates with $H = 0.60$ and $H = 0.90$; (a) and (c) Higuchi estimator, (b) and (d) Peng estimator.

Compared to the other nine estimators implemented in `fSeries` the Whittle estimator was remarkable for its narrow confidence interval. It only displayed a small amount of downwards bias when the series were short and $H$ was high. There was an implementation issue in the software we used. The Whittle estimator would terminate with an error when $H$ was low and the series contained only a few hundred observations. Thus in Figure (7.5)(a) there was no data for series with less than 300 observations in the $H = 0.65$ results.

The results for the wavelet estimator are presented in Figures (7.5) (b) and (d). The wavelet estimator was unbiased for all $H$ values at series lengths over 4,100 data points. The bias present in series shorter than 4,100 data points was very small. Because a discrete wavelet transform can be interpreted as filtering a time series with a series of

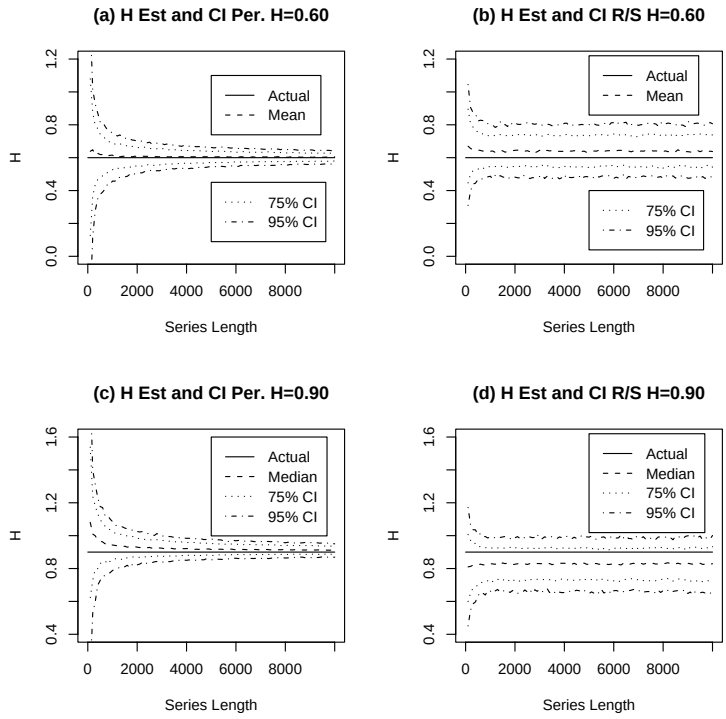


Figure 7.4: Empirical confidence intervals for the $H$ estimates with $H = 0.60$ and $H = 0.90$; (a) and (c) periodogram estimator, (b) and (d) R/S estimator.

*octave* band pass filters, with each doubling of the length of the series a new *octave* of the wavelet becomes available, see McCoy and Walden (1996) for mathematical details. The availability of each new *octave* can be seen in Figures (7.5) (b) and (d) as a series of steps in the reduction of the confidence interval of the estimate with increasing series length. The estimator had constant variance when the number of *octaves* was constant.

The results for the Haslett-Raftery estimator are presented in Figures (7.6)(a) and (b). The Haslett-Raftery did not report estimates of $d$ less than zero ($H < 0.5$). Hence for low $d$ and short series the distribution was truncated on the low side at $d = 0$ or $H = 0.5$ as in Figure (7.6) (a). The Haslett-Raftery estimator was an excellent estimator with only small amounts of bias in the short series and had a narrow confidence interval.

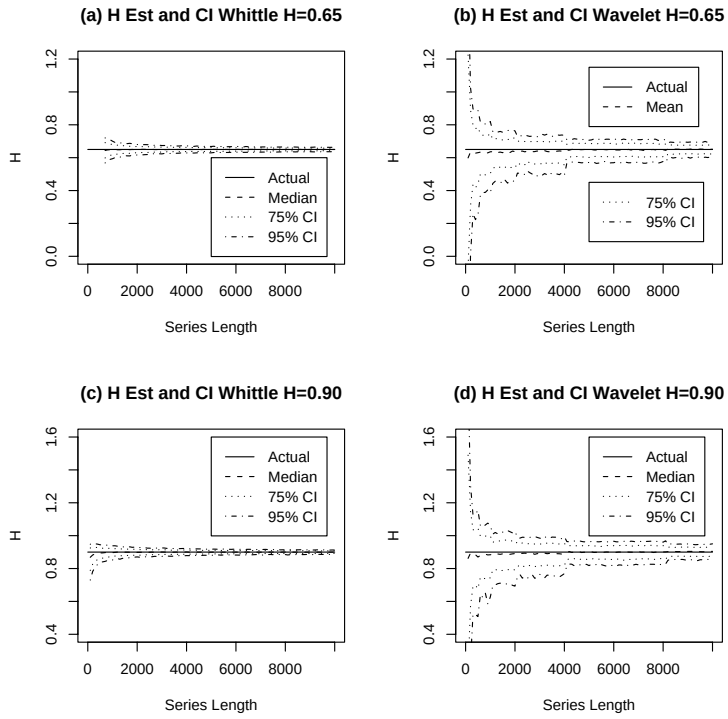


Figure 7.5: Empirical confidence intervals for the $H$ estimates with $H = 0.65$ and $H = 0.90$; (a) and (c) Whittle estimator, (b) and (d) wavelet estimator.

Figure (7.7) presents the MSEs for the estimators for $H = 0.9$ or $d = 0.4$ as appropriate. This is an alternative way to look at the data from the simulations. We only report MSEs for series of 500 data points and longer because of the high MSEs for some estimators in the short series. The Whittle and Haslett-Raftery both had low MSEs in all series greater than 500 data points in length.

## 7.5 Conclusions

Of the eleven estimators examined here the Whittle and Haslett-Raftery estimators performed the best on simulated series. They only exhibited bias in short series with high $H$ or $d$ values and had the lowest and second lowest MSEs respectively of the 11

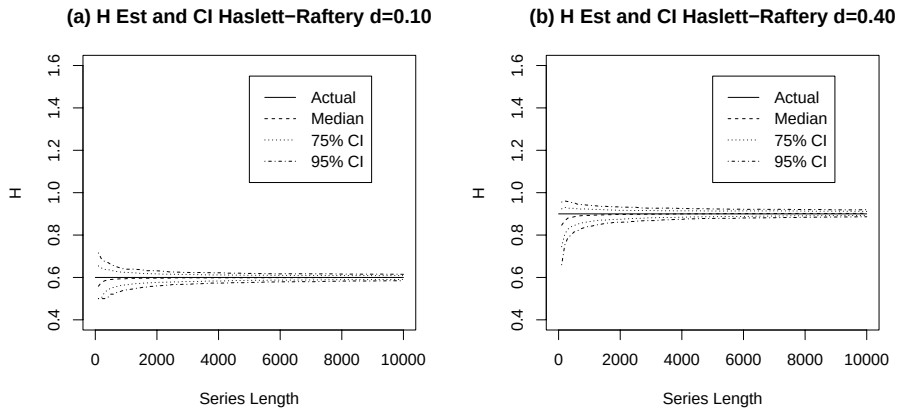


Figure 7.6: Empirical confidence intervals for the $H$ estimates with $d = 0.10$ ($H = 0.60$) and $d = 0.40$ ($H = 0.90$); (a) and (b) Haslett-Raftery estimator.

estimators at all series lengths. These two estimates will be the estimators of choice for the remainder of this thesis. These empirical results are supported for the Whittle estimator by the theoretical work of Fox and Taqqu (1986) and Dalhaus (1989).

For series with 4,000 or more data points, the Peng, periodogram and wavelet estimators look to be good choices based on their MSEs.

The Higuchi estimator is useful if the researcher wishes to recover the fractal dimension of the time series. In contrast to the other estimators it provides useful information on a time series if the series is not an FGN or FI(d) series. As an estimator of $H$ it is inferior to several others.

The boxed periodogram method was clearly inferior to the periodogram method it was intended to improve upon when applied to simulated FGNs. Further research would be needed to test if it is more robust than the periodogram method in series with departures from a pure FGN. This could be accomplished, for example, by simulating ARFIMA series with non-zero AR and MA components or series with structural breaks.

The R/S estimator is of considerable historical interest but had a major deficiency in that its MSE plateaued while all other estimators' MSEs decreased with increasing series length.

The differenced variance estimator was the worst of the eleven estimators in short series. For series longer than 6,000 data points its MSE was better than the R/S and on a par with the absolute value, aggregated variance and Higuchi methods. Teverovsky and Taqqu (1999) did not recommend the use of the differenced variance estimator in isolation as it was part of a test for shifting means or deterministic trends. Teverovsky and Taqqu (1999) also recommended the aggregated and differenced variance plots always be examined visually. We agree with these recommendations. We did not test its robustness to shifting means or deterministic trends. Some numerical results of its

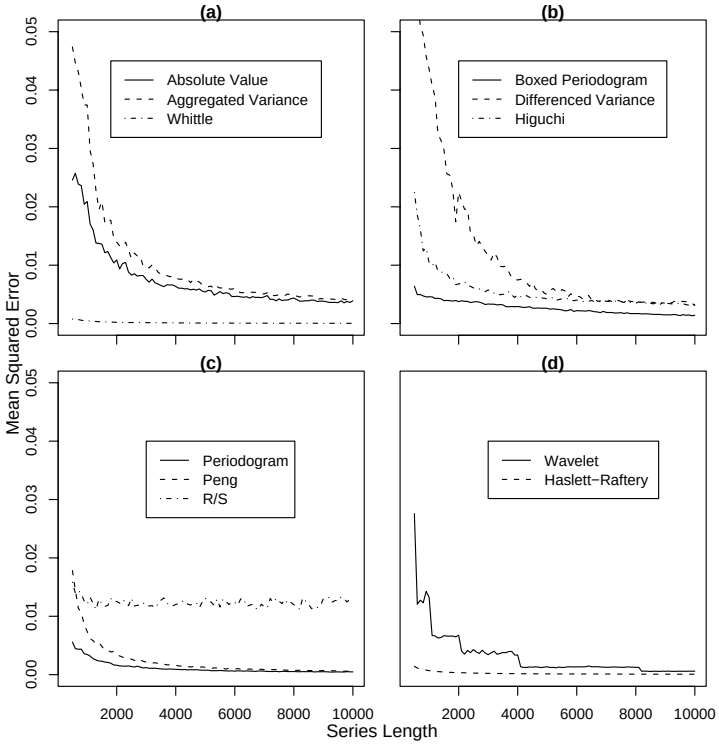


Figure 7.7: Mean squared errors (MSE) as a function of series length for all 11 estimators with $d$=0.4 for the Haslett-Raftery and $H$=0.9 for the other ten. MSEs are reported starting at a series of 500 data points. (a) Absolute Value, Aggregated Variance and Whittle. (b) Boxed Periodogram, Difference Variance and Higuchi. (c) Peng, Periodogram and R/S. (d) Wavelet and Haslett-Raftery.

performance in these two situations can be found in Teverovsky and Taqqu (1999).

# Chapter 8

# Application: Structural Breaks and Long Memory

## 8.1 Introduction

In Chapter (6) we presented a brief summary of a small number of models proposed for long memory time series. Despite the obvious theoretical differences between the FGNs or FI(d)s and constrained non-stationary models, they are difficult, if not impossible, to tell apart in real data. The literature on this problem is growing quickly. Some of the more important papers are Diebold and Inoue (2001), Granger and Hyung (2004), Sibbertsen (2004) and Banerjee and Urga (2005). The difficulty of distinguishing between long memory and structural change was reviewed by Diebold and Inoue (2001) who suggest '... in the sorts of circumstances studied in this paper, "structural change" and "long-memory" are effectively different labels for the same phenomenon ...'.

Sibbertsen (2004) pointed out that the reason distinguishing between long memory and structural breaks is so difficult is because their finite sample properties are similar and so standard methodologies fail. Structural break detection and location techniques, such as ART, report breaks when only long memory is present. Conversely, long memory estimators report long memory when only structural breaks are present.

It is of interest to both theoreticians and practitioners to know the statistical prop-

erties of procedures for detecting and quantifying long memory when only structural change is present. Similarly, it is of interest to know the statistical properties of procedures for detecting and locating structural change when, in fact, there is only long memory. The former problem has received considerable attention while the literature on the latter problem is somewhat sparse, see Sibbertsen (2004) for a survey.

Nevertheless, a number of authors have attempted to develop statistical tests to distinguish between true long memory and other types of processes displaying statistical long memory. Some examples are:

1. Beran (1992) proposed a goodness-of-fit test for FGNs and ARFIMA models by comparing the estimated spectrum for the data with the theoretical spectrum of an FGN or ARFIMA model with the specified parameters. The null hypothesis is that the times series is an FGN or ARFIMA. Rejection of the null does not lead to a clear alternative.

2. Beran and Terrin (1996, 1999) developed a test for which the null hypothesis is that the $H$ parameter is constant in an FGN. As a constant $H$ is fundamental to the definition of an FGN, the Beran and Terrin (1996, 1999) test could be considered to distinguish between an FGN and a multifractal as in multifractals the $H$ (or $d$) parameter is allowed to change over time.

3. Teverovsky and Taqqu (1999) proposed a procedure based on comparing the variance plots from the aggregated variance estimator with the differenced variance estimator which allowed the researcher to distinguish between long memory on the one hand and shifting means or deterministic trends on the other. The null is that the process is an FGN with no level shifts or deterministic trends. The alternative is that either level shifts or deterministic trends or both are present in the data but, the test provides no further information on their number or location.

4. Smith (2005) modified the GPH estimator to discriminate between infrequent level shifts and long memory. The robustified estimator should report an estimate of $d$ close to zero if there are only level shifts in the data.

5. Berkes et al. (2006) presented a procedure for distinguishing between a long memory series and a short memory series with change-points in the mean using a test statistic based on the CUSUM test. It takes the null hypothesis as a short memory series with breaks and the alternative as a long memory series.

6. Ohanissian et al. (2008) developed a test of self-similarity based on different levels of aggregation of the series which cleverly avoids the issue of having to estimate any breaks or additional short memory correlation. The null hypothesis is that $d$ is constant at all levels of aggregation as required by the definition of an FI(d) process, hence the series is a true FI(d) series. Rejection of the null does not lead to a clear alternative.

Our new procedures, presented in Section (8.2) below, add to this literature by approaching the problem from the direction of explicitly modeling the breaks and are motivated by the following considerations.

As mentioned above, the use of structural break detection and location methods are regarded as problematic because they tend to report breaks in FGNs and FI(d) series even though the data generating process is uniform throughout. For example, Wright (1998) proved that when the standard cumulative summation (CUSUM) test, formalized by Brown et al. (1975), for detecting structural breaks was applied to long memory series the probability of reporting a break converged to one with increasing series length. Because it is known that such tests will report breaks with probability one as the series length increases, numerous papers in the literature criticize the use of structural break tests on long memory time series. Examples are Rudnick and Davis (2003) in the environmental literature and Prodan (2003) in the econometric.

We note that if a series is generated by a true long memory process, dividing the series into a number of "regimes" of differing lengths through the use of a structural break location method will only yield subsamples of a single population. However, if, in fact, the series contains structural breaks, the use of a structural break location method will instead divide the series into a number of subpopulations. In the former case our *a priori* expectation is that the subsamples will have the same statistical

properties as the full series. In the latter case the data generating process has one or more discontinuities. Consequently, the statistical properties of the regimes will need to be individually estimated.

Despite this risk of model misspecification we could find no empirical study of the statistical properties of the "regimes" in simulated FGNs or FI(d) series of finite sample size when they were incorrectly analyzed by applying structural break location methods to them. The computational speed of ART has allowed large scale simulation studies to proceed which would have been computationally impractical with established techniques such as that due to Bai and Perron (1998, 2003).

## 8.2 Proposed New Procedures

Motivated by the above considerations we propose two new procedures based on ART to distinguish between true long memory and shifting means: a numerical method which is computationally fast, and a graphical method which is computationally slow but has greater sensitivity. In both procedures the null hypothesis is that the series is an FGN or FI(d) as appropriate. The alternative is that the series has one or more structural breaks.

### 8.2.1 Numerical – ART with Beran (1992) Test

1. Estimate $H$ or $d$ for the full series.
2. Estimate goodness-of-fit of this $H$ or $d$ value with the Beran (1992) test. If the null is not rejected proceed to the next step.
3. Apply ART to the series to obtain the candidate breakpoints.
4. Estimate $H$ or $d$ for each regime.
5. Apply the Beran (1992) test to each regime twice. Once using $H$ or $d$ as estimated for the full series and once using $H$ or $d$ as estimated for the regime.

6. Assess whether $H$ or $d$ is constant across series. Details of how the assessment was made are given in Section (9.3).

### 8.2.2 Graphical – Using ART to Estimate the Bivariate Distribution of $H$ or $d$ with Regime Length.

1. Estimate $H$ or $d$ for the full series.
2. Estimate goodness-of-fit of this $H$ or $d$ value with the Beran (1992) test.
3. Through simulation obtain the bivariate probability distribution of $H$ or $d$ with regime length.
   (a) Simulate a large number of FGNs or FI(d) series with the same $H$ or $d$ value and series length as the series under test.
   (b) Use ART to break the simulated series into "regimes".
   (c) Estimate $H$ or $d$ within these regimes.
   (d) Calculate the empirical median, 75%, 95%, and 99% confidence intervals.
   (e) Plot the bivariate distribution of $H$ or $d$ against regime length.
4. Apply ART to the full series to obtain the candidate breakpoints.
5. Estimate $H$ or $d$ for each regime.
6. For each regime overplot the $H$ or $d$ and regime length on the previously determined empirical distribution.
7. Assess whether $H$ or $d$ is constant. Details of how the assessment was made are given below in Section (9.3.4) .

In step 2 in the above methods the question remains what to do if the Beran test rejects the null. This depends on the researcher's goals. If they are only concerned with estimating the goodness-of-fit then their task may be complete and there may not be much point proceeding to a long set of simulations to perform the graphical test. If

the researcher wishes to find the reasons for the particular data set's departure from an ideal FGN or FI(d) process then they may still wish to perform the graphical test.

In step 3a of the Graphical method we found that typically it was necessary to simulate 10,000 to 25,000 FGNs or FI(d) series to obtain a good estimate of the bivariate distribution. The graphical procedure can also be used to assess whether the distributions of the mean, standard deviation, skewness, and kurtosis are consistent with an FGN or FI(d) series. We can use a very similar method to obtain the empirical bivariate distribution of number of breaks against CUSUM range. This bivariate breaks-CUSUM distribution is used in Sections (9.3.7) and (10.6.5) below.

## 8.3 Distribution of Breaks in an FI(d) Series

In Chapter (3) we examined ART's performance in detecting structural breaks in a range of short memory time series. In this section we examine ART's behaviour when applied to simulated long memory time series. Thus we undertake a simulation study of the distribution of the numbers of breaks, the break locations and regime lengths reported by ART for simulated FI(d) series. This was motivated by the observation of Nunes et al. (1995) who noted in simulation experiments that when the error terms in a regression model follow an I(1) process there was a tendency for structural break estimators to estimate a breakpoint in the middle of the sample when, in fact, the breakpoint did not exist. Subsequently, Bai (1998) provided a mathematical proof for this phenomenon.

We simulated FI(d) series with lengths from 1,000 to 16,000 data points in steps of 1,000 data points, and $d$ values between 0.02 and 0.48 in steps of 0.02. ART was applied to each series and the number of breaks, their locations and associated regime lengths were recorded. For each set of parameter values 1,000 replications were run.

The distribution of the number of breaks for series with 2,000 and 8,000 data points is presented in Figure (8.1), the remainder are omitted for reasons of space.

It is clear from the graphs that as the value of $d$ increased the distribution made a transition from all or almost all series having no reported breaks to all series having

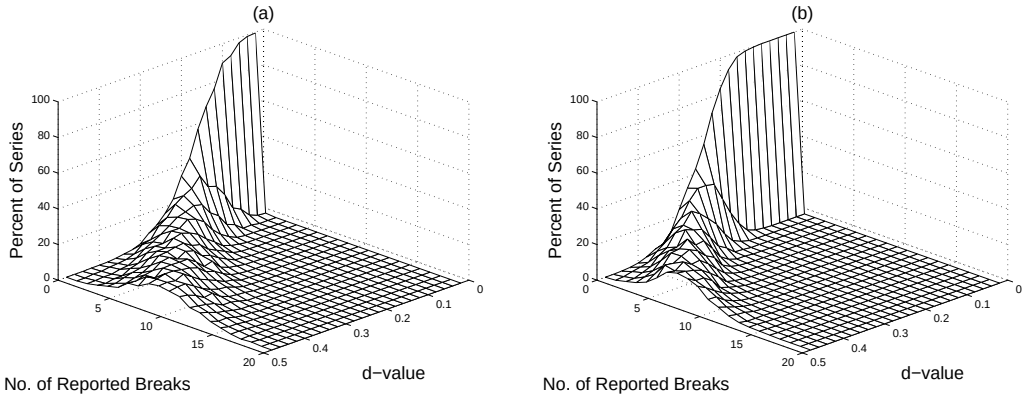


Figure 8.1: Distribution of number of breaks reported by ART for 1000 replications of different values of $d$ in an FI(d) series. Panel (a) 2000 and panel (b) 8000 data points.

multiple reported breaks. It is precisely this phenomena which has made it so difficult to distinguish between true long memory and constrained non-stationary series.

It was also clear from the graphs that as the length of the series increased the tendency for ART to report breaks decreased. This observation should be unsurprising as the data generating process was uniform throughout, the process was stationary and the tree pruning method imposed a penalty on numbers of terminal nodes. Asymptotically ART may well report no breaks.

If we fix both the series length and the $d$ value the resulting distribution of the number of breakpoints per series is well approximated by a Poisson distribution[1]. The Poisson distribution has the elegant feature that, numerically, $\mu = \sigma^2 = \lambda$. Because of this property we can obtain some estimate of goodness-of-fit by plotting $(\mu - \sigma^2)/\mu$ against $d$ for the different values of $d$ and series lengths. This is presented in Figure (8.2). As can be seen the fit to the Poisson distribution improved with series length.

Tables (8.1) and (8.2) present the mean number of breaks per simulated series for a range of parameter values. Intermediate values can be calculated by linear interpolation. The same data is presented in graphical form in Figure (8.3).

[1]This was determined by G. Y. Zhang, a research student whom Marco Reale and William Rea co-supervised in the summer of 2006/7. This research was presented in Zhang (2007), which accidentally omitted both Marco Reale and William Rea as co-authors.

Table 8.1: Table of the mean number of breaks for a given $d$-value for series lengths 1000 to 8000.

| $d$ | 1K | 2K | 3K | 4K | 5K | 6K | 7K | 8K |
|---|---|---|---|---|---|---|---|---|
| 0.02 | 0.15 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.04 | 0.23 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.06 | 0.35 | 0.03 | 0.01 | 0 | 0 | 0 | 0 | 0 |
| 0.08 | 0.54 | 0.09 | 0.02 | 0.002 | 0.001 | 0 | 0 | 0 |
| 0.10 | 0.77 | 0.12 | 0.03 | 0.01 | 0.01 | 0.001 | 0.002 | 0 |
| 0.12 | 1.08 | 0.28 | 0.10 | 0.03 | 0.01 | 0.004 | 0.002 | 0.003 |
| 0.14 | 1.47 | 0.40 | 0.17 | 0.10 | 0.05 | 0.03 | 0.02 | 0.01 |
| 0.16 | 1.82 | 0.64 | 0.33 | 0.20 | 0.11 | 0.07 | 0.04 | 0.03 |
| 0.18 | 2.38 | 0.91 | 0.54 | 0.32 | 0.23 | 0.16 | 0.11 | 0.08 |
| 0.20 | 2.86 | 1.26 | 0.75 | 0.47 | 0.37 | 0.28 | 0.22 | 0.19 |
| 0.22 | 3.58 | 1.74 | 1.11 | 0.80 | 0.53 | 0.43 | 0.39 | 0.34 |
| 0.24 | 4.15 | 2.19 | 1.44 | 1.15 | 0.82 | 0.71 | 0.64 | 0.57 |
| 0.26 | 4.81 | 2.71 | 2.00 | 1.56 | 1.27 | 1.06 | 0.94 | 0.85 |
| 0.28 | 5.60 | 3.24 | 2.48 | 1.95 | 1.60 | 1.42 | 1.22 | 1.11 |
| 0.30 | 6.14 | 4.03 | 3.11 | 2.55 | 2.19 | 1.92 | 1.71 | 1.54 |
| 0.32 | 6.94 | 4.61 | 3.72 | 3.17 | 2.76 | 2.56 | 2.27 | 2.16 |
| 0.34 | 7.58 | 5.24 | 4.41 | 3.76 | 3.41 | 3.16 | 2.90 | 2.75 |
| 0.36 | 7.86 | 5.93 | 5.00 | 4.50 | 3.97 | 3.69 | 3.51 | 3.24 |
| 0.38 | 8.54 | 6.62 | 5.80 | 5.12 | 4.77 | 4.46 | 4.24 | 4.04 |
| 0.40 | 9.12 | 7.25 | 6.37 | 5.92 | 5.59 | 5.27 | 4.95 | 4.68 |
| 0.42 | 9.52 | 7.98 | 7.05 | 6.60 | 6.29 | 5.88 | 5.69 | 5.48 |
| 0.44 | 9.85 | 8.48 | 7.50 | 7.22 | 6.87 | 6.56 | 6.39 | 6.27 |
| 0.46 | 10.14 | 8.91 | 8.23 | 7.83 | 7.48 | 7.12 | 7.03 | 6.90 |
| 0.48 | 10.05 | 9.21 | 8.72 | 8.34 | 7.98 | 7.85 | 7.65 | 7.46 |

Table 8.2: Table of the mean number of breaks for a given $d$-value for series lengths 9000 to 16000.

| $d$ | 9K | 10K | 11K | 12K | 13K | 14K | 15K | 16K |
|---|---|---|---|---|---|---|---|---|
| 0.02 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.04 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.06 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.08 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.14 | 0.01 | 0.006 | 0.003 | 0.001 | 0.001 | 0.001 | 0 | 0 |
| 0.16 | 0.02 | 0.01 | 0.005 | 0.004 | 0.002 | 0.002 | 0.002 | 0.005 |
| 0.18 | 0.07 | 0.04 | 0.05 | 0.03 | 0.03 | 0.03 | 0.02 | 0.009 |
| 0.20 | 0.15 | 0.11 | 0.08 | 0.07 | 0.06 | 0.05 | 0.04 | 0.04 |
| 0.22 | 0.27 | 0.24 | 0.21 | 0.18 | 0.15 | 0.14 | 0.13 | 0.12 |
| 0.24 | 0.50 | 0.43 | 0.37 | 0.32 | 0.29 | 0.27 | 0.24 | 0.24 |
| 0.26 | 0.74 | 0.68 | 0.60 | 0.55 | 0.51 | 0.47 | 0.45 | 0.45 |
| 0.28 | 1.07 | 0.99 | 0.94 | 0.90 | 0.81 | 0.75 | 0.70 | 0.68 |
| 0.30 | 1.45 | 1.39 | 1.33 | 1.27 | 1.20 | 1.1 | 1.06 | 1.02 |
| 0.32 | 1.97 | 1.89 | 1.80 | 1.72 | 1.67 | 1.59 | 1.49 | 1.34 |
| 0.34 | 2.55 | 2.42 | 2.3 | 2.22 | 2.16 | 2.09 | 2.06 | 1.96 |
| 0.36 | 3.24 | 3.10 | 2.94 | 2.83 | 2.64 | 2.57 | 2.52 | 2.43 |
| 0.38 | 3.79 | 3.58 | 3.53 | 3.49 | 3.32 | 3.29 | 3.18 | 3.17 |
| 0.40 | 4.63 | 4.48 | 4.29 | 4.17 | 4.09 | 4.01 | 3.93 | 3.83 |
| 0.42 | 5.35 | 5.12 | 5.07 | 4.95 | 4.8 | 4.72 | 4.75 | 4.60 |
| 0.44 | 6.16 | 6.06 | 5.97 | 5.84 | 5.70 | 5.64 | 5.58 | 5.30 |
| 0.46 | 6.81 | 6.72 | 6.51 | 6.46 | 6.30 | 6.28 | 6.09 | 6.16 |
| 0.48 | 7.34 | 7.31 | 7.19 | 7.13 | 7.08 | 6.99 | 6.88 | 6.86 |

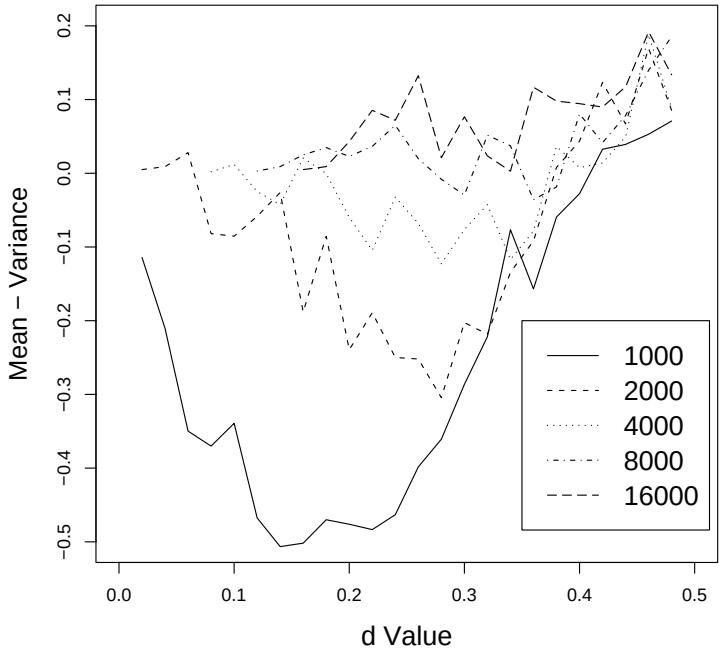


Figure 8.2: Plot of $(\mu - \sigma^2)/\mu$ against $d$ for series lengths ranging from 1,000 to 16,000 data points where $\mu$ is the mean number of breaks reported by ART in and FI(d) series and $\sigma^2$ is the estimated variance of the numbers of breaks.

A crude but very simple test of long memory vs structural breaks arises from these observations. This test is used in Zhang (2007) who indicated the test was only effective when $0.8 < H < 1.0$. While Zhang (2007) did not consider FI(d) processes the test should also be effective for $0.3 < d < 0.5$.

1. Estimate $H$ or $d$ for the full series.

2. Apply the Beran (1992) test to the full series to assess whether it is likely to be an FGN or FI(d) series.

3. Use Tables (8.1) and (8.2) to find the expected number of breaks for that series length and $H$ or $d$ value.

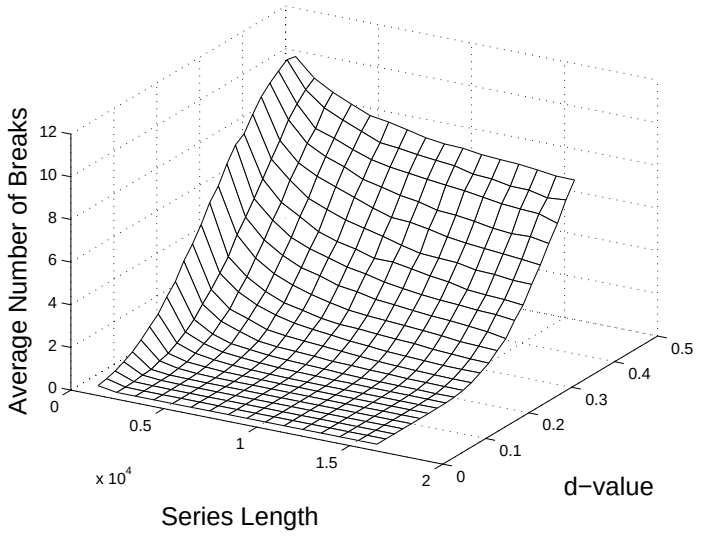


Figure 8.3: Mean number of breaks reported by ART for different values of $d$ in an FI(d) series with lengths ranging from 1000 to 16000.

4. Apply ART to the series and obtain the number of reported breaks.

5. Use the Poisson distribution to assign a p-value to the probability that the series is a true long memory process.

### 8.3.1 Distribution of Lengths of Regimes

An alternative to looking at the number of breaks in simulated series is to consider the lengths of the regimes. Figure (8.4) presents the results for simulated FI(d) series with 4000 data points, the remainder are omitted for reasons of space.

When comparing Figures (8.1) and (8.4) it should be clear that when breaks were reported there was a tendency to report regimes which were quite short in comparison with the whole series. The change of behaviour of ART between values of $d$ for which ART reported no breaks and values for which breaks were reported suggested that to distinguish between long memory and regime switching processes at least two approaches were required. Tests or procedures involving ART would be useful when

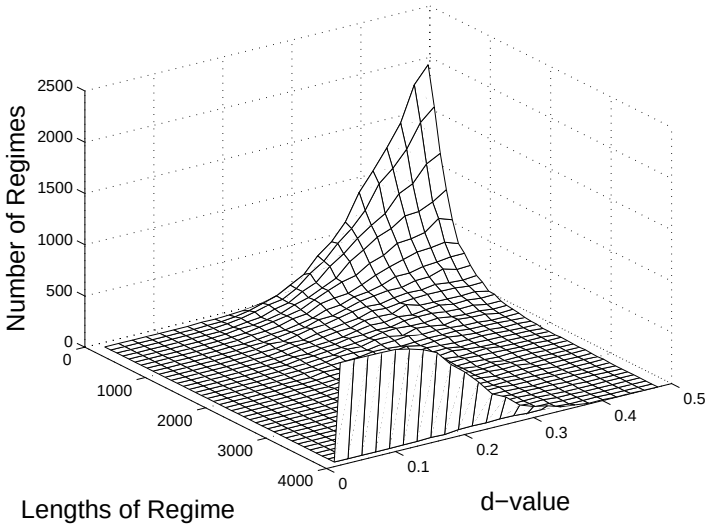


Figure 8.4: Distribution of regime lengths reported by ART for 1000 replications of different values of $d$ in simulated FI(d) series of 4000 data points.

$H$ or $d$ was sufficiently high that a reasonable number of regimes would expected to be reported. When $H$ or $d$ was sufficiently low that no breaks would be expected to be reported some alternative method would need to be used, with the exception that if no breaks are expected and at least one is reported then it is likely that the break is real. For financial data with a typical $d$ value of about 0.40 and several thousand observations ART should be useful.

### 8.3.2 Break Locations

In this subsection we examine the distribution of the break locations for simulated FGNs with known $H$ values. Figure (8.5) presents the break locations for six values of $H$ in simulated FGN series 4000 data points in length. The remainder of the results are omitted for reasons of space. A variable number of replications were run because when $H$ was low few breaks were reported. Thus we required a much larger number of replications than when $H$ was high to obtain a reasonable estimate of the break

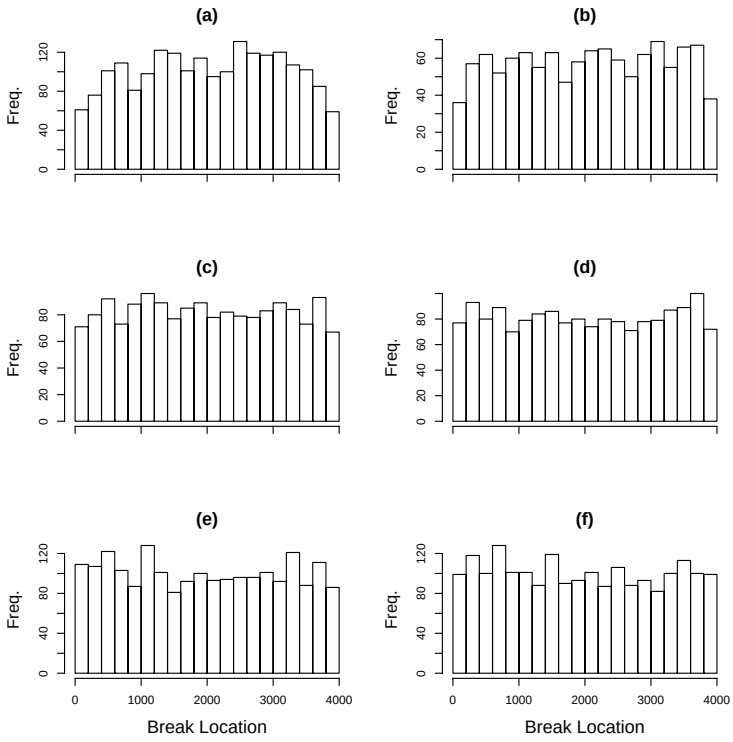


Figure 8.5: Distribution the locations of the breaks reported by ART for different values of $H$ in simulated FGN series of 4000 data points. (a) $H = 0.60$, (b) $H = 0.65$, (c) $H = 0.70$, (d) $H = 0.75$, (e) $H = 0.80$, (f) $H = 0.85$.

location distribution. It seems reasonable to believe these were uniformly spread across the series. There was an exception for the series with $H = 0.60$ or $H = 0.65$ but this was probably due to the fact that, on average, there was less than one break per series and so there were few breaks reported near the end points of the series.

## 8.4 Bivariate Distributions of $d$ and Regime Length

In this section we examine the bivariate distributions of $H$ or $d$ estimates and regime length as required for the graphical method presented in Section (8.2.2) above for four types of constrained series and provide a comparison with simulated FI(0.40) series.

Technically these series are stationary but they, or similar series, have been used in the literature as proxies for more general non-stationary series. The choice of $d = 0.40$ was guided by the observation that financial time series often exhibit long memory with $d = 0.40$. See the literature review and case study in Chapter (9) for further details. The four constrained non-stationary series were:

**Markov Regime Switching:** In these models the series had two states and switched between them according to a probabilistic rule. They have been studied in the past by a number of authors. See for example Granger and Teräsvirta (1999) and Smith (2005). We used a slightly different model in that we fixed the number of switches per series and then determined the location of the switches using a uniform distribution. We placed two constraints on the location of the state changes. We did not allow the state to change within 50 data points of either end of the series and the series was required to remain in a state for at least 20 data points before a switch was permitted.

**Uniform Regime Switching:** These models differed from the previous one by generating a new mean after a switch by a uniform distribution rather than switching between fixed states.

**Constrained Wandering Mean:** In these models the range of the mean was constrained to lie between a minimum and maximum value but was not stationary. At each break a new rate and direction of drift was selected. Further, if the mean reached one of the allowable extreme values it reversed direction, but retained the same magnitude of drift until the next break.

**Multifractals:** In these models at each break a new $H$ value and mean was chosen. Within each regime the data was an FGN.

With the exception of the simulated multifractals, simple tests such as that of Beran (1992) were adequate to determine the simulated Markov, structural break and constrained wandering mean series were not FGNs or ARFIMA series. While detailed

results are not presented here the Beran test did not accept the null hypothesis at the five percent level in either 100 percent or close to 100 percent of the simulated series in 1000 replications. For multifractals the Beran test did not accept the null of an FI(d) series in about 84 percent of seven break simulated multifractal series at the five percent level. The rejection rate of the null at the five percent level ranged from about 73 percent in the five break series to 91 percent in the 25 break multifractal.

In the individual papers cited in Section (8.1) there are usually small simulation studies to show that their proposed tests performed well by reporting high rates of rejection of the null hypothesis on the types of series described in this section.

In the simulation study below we used the Haslett-Raftery estimator rather than the Whittle because the Whittle implementation often exited with an error when applied to regimes of less than 100 data points.

### 8.4.1 Simulated FI(0.40) Series

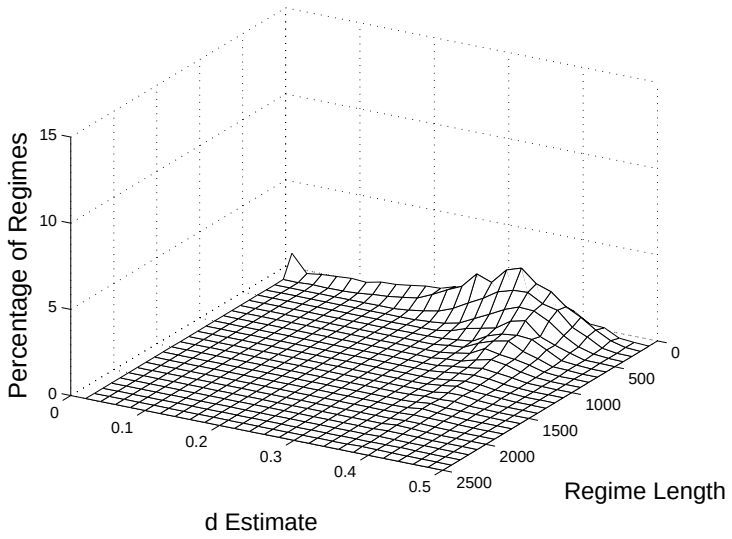


Figure 8.6: Empirical bivariate distribution the estimates of $d$ and regime length for an FI(d) series with d=0.40 and a length of 2500 data points.

We ran 25,000 replications of simulated FI(0.40) series with length 2500 data points. The empirically determined bivariate distribution is presented in Figure (8.6). The two parameters, the $d$ estimate and the regime length, are plotted on the two horizontal axes and the density is plotted on the vertical axis.

For the simulated FI(0.40) series, splitting the series up into "regimes" with ART did not mask the fact that the series had long memory. It was only in the very short regimes that the Haslett-Raftery estimator returned any values close to zero. We recall from Chapter (7) that the Haslett-Raftery estimator exhibited a small amount of downward bias when the series was short and the $d$ value high. That bias can been in Figure (8.6) as the $d$ estimates reported in the short regimes were lower than the nominal value of 0.40.

### 8.4.2 Markov Switching Models

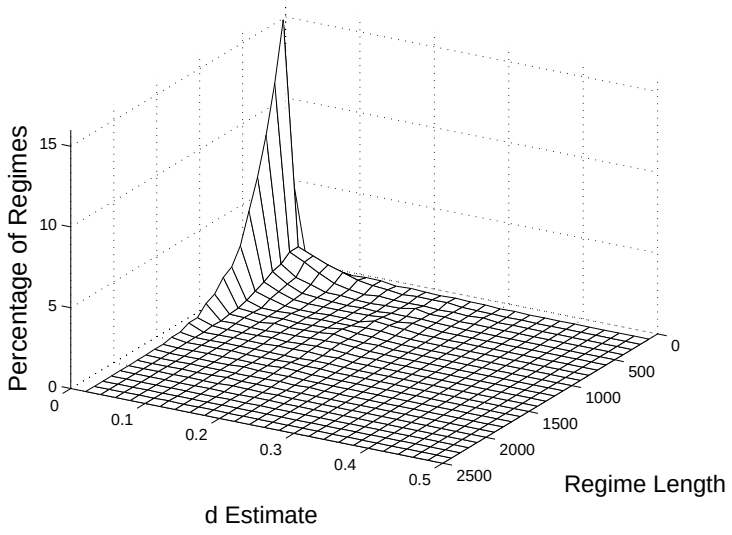


Figure 8.7: Empirical bivariate distribution of $d$ estimates with regime length for a Markov switching model with seven breaks and a length of 2500 data points.

We examined Markov switching models with 2500 data points and five through 12,

16, 19, 22, and 25 breaks. From Table (8.1) we can see that for a series of this length and $H = 0.90$ ($d = 0.40$) that the expected number of breaks is about seven. Thus we report the results for the seven break simulations. The remainder are omitted for reasons of space.

Although we do not report detailed results, for reasons of space, the number of breaks had only a small influence on the estimates of $H$ or $d$ reported by the Whittle or Haslett-Raftery estimators. For five breaks the $H$ estimates lay in the range 0.66 to 0.78, while for 25 breaks the range was 0.74 to 0.79. The Haslett-Raftery tended to report slightly higher $d$ estimates than their $H$ equivalents. For the seven break series the Haslett-Raftery reported a mean of $d = 0.29$

Thus the simple Markov switching model generated series with $H$ or $d$ estimates lower than those typically seen in geophysical and financial data sets.

In Figure (8.7) we present a three-dimensional surface plot of the empirical bivariate probability density function of Markov switching series with seven breaks.

It is clear that ART had correctly located the switchpoints and hence divided the series into regimes which consisted of nothing more than random numbers with the same mean. The Haslett-Raftery estimator was then applied to these regimes and it correctly reported a $d$ value of zero or close to zero.

### 8.4.3 Uniform Switching Models

As with the simulated Markov switching series we examined structural break switching models with five through 12, 16, 19, 22, and 25 breaks. For the same reasons we only report the results for the seven break simulations.

The Whittle and Haslett-Raftery estimators reported very similar results to those for the simulated Markov switching series.

In Figure (8.8) we present a three-dimensional surface plot of the empirical bivariate probability density function of Uniform switching series with seven breaks.

It is clear that ART had, in general, correctly located the switchpoints and hence divided the series into regimes consisting of nothing more than random numbers with

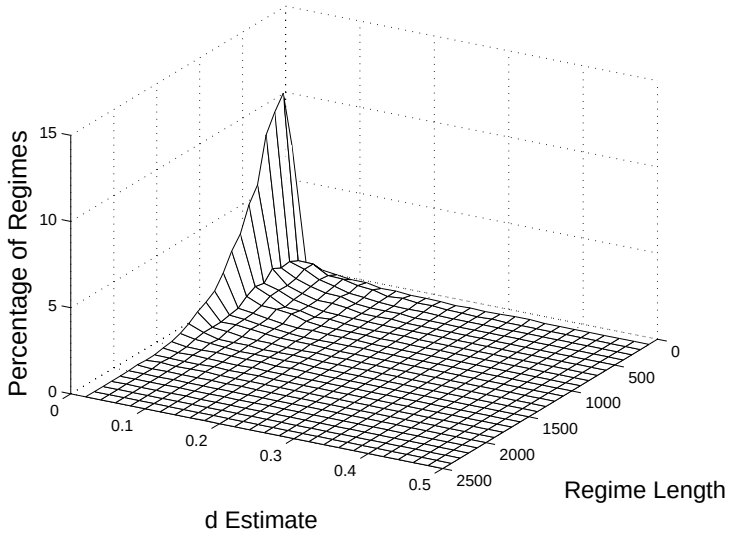


Figure 8.8: Empirical bivariate distribution of d estimates with regime length for Uniform switching models with seven breaks and a length of 2500 data points.

the same mean. The Haslett-Raftery estimator was then applied to these regimes and it correctly reported a $d$ value of zero or close to zero.

### 8.4.4 Wandering Mean Models

As with the two previous simulated series we examined wandering mean models with five through 12, 16, 19, 22, and 25 breaks. For the same reasons we only report the results for the seven break simulations.

The Whittle and Haslett-Raftery estimators reported slightly lower $H$ and $d$ estimates for the full series than those reported for the Markov and Uniform switching series. For the five break series the Whittle reported estimates ranging from $H = 0.58$ to $H = 0.68$ while it reported estimates ranging from $H = 0.52$ to $H = 0.68$ for the 25 break series.

This was a type of series which ART was not tested against in the simulations in Chapter (3). The results are presented in Figure (8.9). While these results are similar

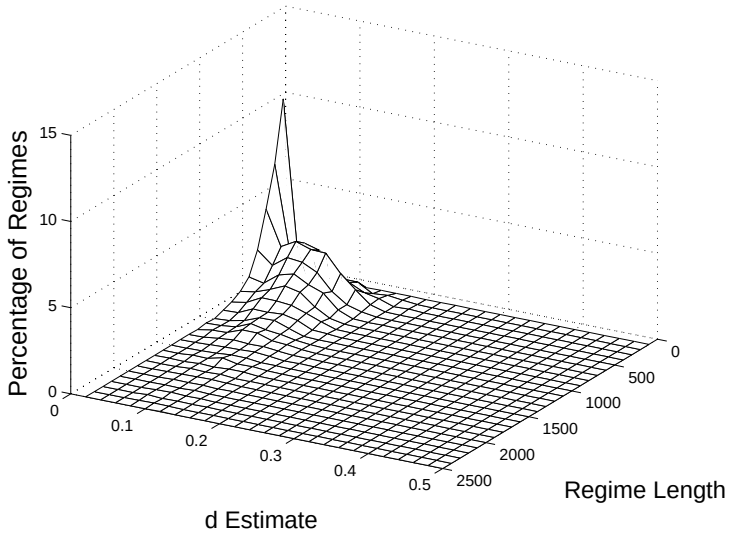


Figure 8.9: Empirical bivariate distribution of $d$ estimates with regime length for Wandering Mean models with seven breaks and a length of 2500 data points.

to those presented in Figures (8.7) and (8.8) it appears that ART had missed a number of breaks so that the bivariate $d$ and regime length distribution did have a region where the $d$ estimates were between zero and 0.20.

### 8.4.5 Multifractal Models

There are many ways to generate a multifractal. In the series reported here we fixed the number of breaks and generated their locations by a uniform distributions. The means of the regimes switched between plus and minus one as with the Markov switching series above. At each break at new $d$ value between 0.15 and 0.30 was generated using a uniform distribution and FI(d) series of the length of the regime were generated. As with the previous simulations we simulated series with five through 12, 16, 19, 22, and 25 breaks. Again we only report on the seven break multifractals for reasons of space.

The Whittle estimator reported a mean of $H = 0.83$ while the Haslett-Raftery estimator reported a mean of $d = 0.39$ for 1000 replications of this type of multifractal

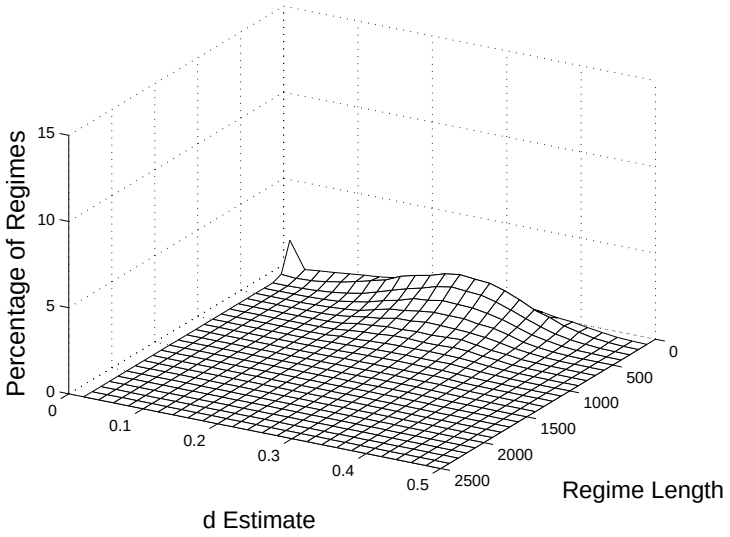


Figure 8.10: Empirical bivariate distribution of $d$ estimates with regime length for Multifractal models with seven breaks.

series. Recall from Section (8.4) that the Beran test exhibited less power against the multifractals than the three previous regime switching type models.

The empirical bivariate distribution for this particular type of multifractal is presented in Figure (8.10). As can be seen very little of the density of this distribution lay in the parameter region above $d = 0.39$. Thus the $d$ estimates within the regimes as reported by ART are, in general, lower than the $d$ estimate for the full series.

Barely visible on the back right hand side of Figure (8.10) is the fact that probability density declined as there were few regimes with less than 100 data points. This differed from the distribution of a simulated FI(d) series such as that in Figure (8.6) in two ways. The first was that the probability density went to zero for regime lengths longer than approximately 1,500 data points. Secondly the distribution in the short regime regions was wider than for FI(d) series.

Table 8.3: Percentages of "regimes" for which the Beran (1992) test yielded an improvement in fit if the regime estimate of $H$ was used instead of the series value in simulated FGNs.

| Result | Percentage | Mean Change |
|---|---|---|
| Improvement | 63.8 | 0.068 |
| Same or Worse | 36.2 | -0.079 |

### 8.4.6 Empirical Testing of Beran Test

In the case studies we utilized the Beran test both alone and in conjuction with ART. We performed a small simulation study of the Beran test to determine for what proportion of tests it reported a better fit if we tested against the within regime $H$ estimate rather than the series $H$ estimate. We performed 1000 replications of a 2500 data point series with $H = 0.90$. The results are presented in Table (8.3).

## 8.5 Discussion

A number of claims have been made in the literature that regime switching and long memory processes are indistinguishable in finite series. If the data generating processes are as simple as the ones we have simulated here then the bivariate distributions of $d$ with regime length are quite different for the two classes of models. Examining figures (8.6), (8.7), (8.8), (8.9), and (8.10), it would seem we could potentially tell the difference between a series which was either an FGN, FI(d) or a multifractal against an alternative that the DGP was some sort of regime switching process on the basis of comparing the distributions of estimated $H$ or $d$ values and regime lengths for real series against the empirical distributions obtained through simulation. If the bivariate distribution for the series under test suggested we were dealing something resembling an FGN, FI(d) or multifractal, the fact that the Beran test is particularly powerful against underestimation of $H$ or $d$ the method outlined in Section (8.2.1) could potentially

distinguish between FGN or FI(d) process against a multifractal alternative.

## 8.6 Publications and Presentations

Material from this chapter is contained in Zhang (2007).

# Chapter 9

# Stock Market Volatilities

## 9.1 Introduction

Statistical long memory has been observed in financial data sets for many years. Whether this long memory is real or spurious has major implications for financial markets.

A number of studies such as those of Andersen et al. (2001), Martens (2002), and Thomakos and Wang (2003) all highlight that the typical financial asset has realised volatilities that are fractionally integrated of order around 0.4 giving the data a long memory property. In estimating stock market volatilities, which we are considering in this chapter, Bollerslev and Wright (2000) argue that estimates of the degree of fractional integration are unbiased for daily volatility based on intraday returns, whereas they are severely downward biased when estimated from daily squared returns.

Diebold and Inoue (2001) showed that occasional structural breaks could spuriously suggest the presence of long memory. As financial volatility data seem to have occasional, irregular, level shifts, some of which can be identified with historical events, it would seem important to detect and model these shifts when trying to establish the degree of long memory (and vice versa). On the other hand, using S&P 500 realised volatility data, Martens et al. (2004) reported that level shifts did not account for the long memory characteristics in the data. The fractional integration parameter declined

when structural breaks were explicitly modeled, but remained significantly greater than zero.

This last statement should not surprise us. For comparatively short series estimators of long memory will report estimates of $H$ or $d$ in the long-memory regions when, in fact, they can be adequately modeled with a low-order short memory ARMA model.

There are a growing number of papers in the econometric literature which suggest that the long-memory is spurious. Mikosch and Starica (2003) suggested that the long memory was simply due to non-stationarity in financial series, and in particular that the mean was not constant. Martens et al. (2004) considered models which included asymmetrical, day of the week, leverage, lagged returns, structural change, and exogenous effects. The models studied by Martens et al. (2004) outperformed simple ARFI models in forecasting. Scharth and Medeiros (2007) studied 16 realised volatility series and concluded that their models, which included smooth transition regression with asymmetrical effects in addition to those studied by Martens et al. (2004), outperformed an ARFIMA model for 15 of the 16 series.

Nevertheless, these studies sometimes showed that ARFIMA models, often with AR and MA orders of zero, outperformed these complicated models when the forecasting horizon extended more than a few days.

While the economics of long memory are beyond the scope of this thesis a few words are in order on the relevance of the long memory vs structural break models to economics and finance[1].

The economics of why highly efficient financial markets would exhibit long range dependence is rarely, if ever, addressed in the literature. Possibly this is because it seems the econometric evidence is overwhelmingly in favor of long memory or, as Diebold and Inoue (2001) concluded, that long memory and short memory with breaks are just two different labels for the same phenomena. Some authors remain unconvinced of their

[1]The material forming the remainder of this section is drawn from section four of W. Rea, L. Oxley, M. Reale, and E. Mendes; Long memory or shifting means? A new approach and application to realised volatility; in preparation for submission. Section four was largely written by L. Oxley with some input from W. Rea.

equivalence, ourselves among them. Mikosch and Starica (2003, p456) concluded "we have tried hard to find in the literature any convincing rational/economic argument in favor of long range dependent stationary log returns, but did not find any." Ohanissian et al. (2008) also state, "the introduction of fractional components into volatility models is purely from an empirical perspective without any theoretical justification."

Some authors have examined the pricing of options and other derivatives under long memory and various short memory assumptions. For example, Bollerslev and Mikkelson (1996) and Taylor (2000) both found there were considerable differences in pricing under the different assumptions. Ohanissian et al. (2008) concluded "ignoring or mis-specifying the nature of observed long memory characteristics in volatility leads to serious option mis-pricing. Specifically when the true DGP is spurious long memory, using either a no long memory model or a true long memory model leads to general under-pricing of call options, by as much as two-thirds. On the other hand, when the DGP is true long memory, using either a no long memory or a spurious long memory leads in general to over-pricing of call options."

In short, the investors, hedgers and speculators trading in the financial markets do not price options as if the long memory is real. Highly efficient financial markets would not tolerate the magnitude of mis-pricing suggested above. Market participants would simply adjust the prices until they were fair. Thus the financial evidence from the markets themselves is that the long memory property exhibited by these series is genuinely spurious and finance theory tells us the issue of which model is "right" is important. A consequence of this is that tests and procedures such those outlined in Section (8.1) ought to report that financial time series are not true long memory processes.

In fact, the opposite is the case. Both Smith (2005) and Ohanissian et al. (2008), who applied their methods to soybean prices and exchange rates respectively, concluded these series were generated by true fractionally integrated processes. Both methods appear soundly based in theory. The reasons for their failure, if indeed it is a failure, to correctly determine the nature of the DGP is unclear.

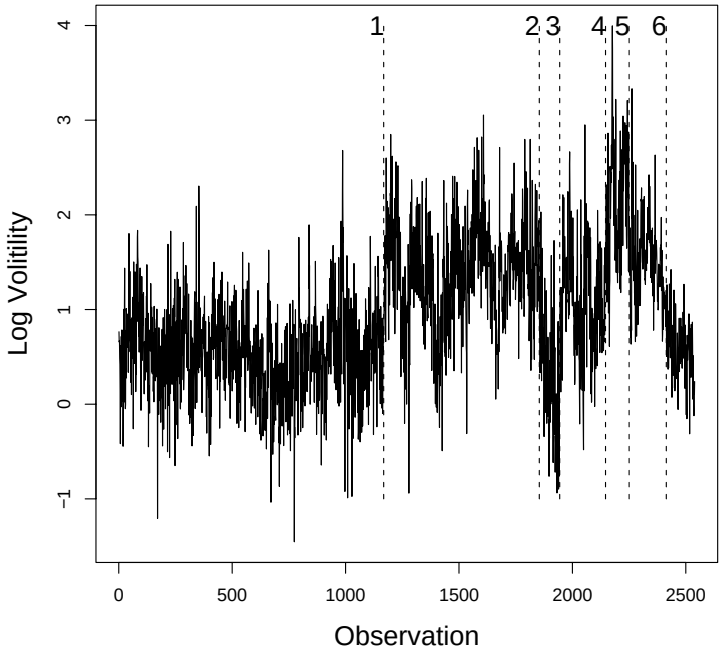


Figure 9.1: AIG realized volatility series plot with breaks reported by ART marked by vertical dashed lines and labeled 1 through 6.

As the financial and econometric literature almost exclusively uses FI(d) models rather than FGNs, in this chapter we will report results in term of $d$ rather than $H$ estimates and unless otherwise stated we used the Haslett-Raftery estimator.

## 9.2 The Data Set

Scharth and Medeiros (2007) kindly provided the data used in this chapter. The data set comprised realized volatility and returns of 14 stocks which trade on the New York Stock Exchange and two which trade on the NASDAQ; Alcoa (AA), American International Group (AIG), Boeing (BA), Caterpillar (CAT), General Electric (GE), Hewlett Packard (HP), IBM, Intel (INTC), Johnson and Johnson (JNJ), Coca-Cola

(KO), Merck (MRK), Pfizer (PFE), Walmart (WMT), and Exxon (XON). The period of analysis was from January 3, 1994 to December 31, 2003. Trading days with abnormally small trading volumes were excluded by Scharth and Medeiros, leaving a total of 2539 daily observations. The daily realized volatility was estimated using the two time scale estimator of Zhang et al. (2005) with five-minute grids, which is a consistent estimator of the daily volatility. A broader explanation of the dataset can be found in Scharth and Medeiros (2007). It should be noted that because all 16 were part of the Dow Jones Industrial Average they cannot be considered to be independent series.

We will give detailed results for American International Group (AIG) and, where practical, state the results for the other 15 stocks.

## 9.3 Results

In this section we present the results of the data analysis. The results of our new procedures are presented in Sections (9.3.3) and (9.3.4) below.

### 9.3.1 Basic Statistics

Andersen et al. (2001) studied the distributions of realized stock market volatility. They found that the distribution was highly right skewed and with heavy tails. However, a log transformation resulted in an approximately normal distribution for the 60 series they studied.

We estimated the mean, standard deviation, skewness and kurtosis for each series and regime within the series as well as applying the Jarque-Bera test for normality. Table (9.1) presents the results for AIG. These types of results were typical of the realized volatility series. All 16 stocks showed this type of behaviour where some regimes showed severe departure from Gaussianity and other regimes showed an excellent fit to the normal distribution.

Figure (9.1) presents a plot of the data with the breakpoints reported by ART marked as vertical dashed lines and labelled one through six. Figure (9.2) presents the

Table 9.1: Basic statistics for the regimes within the AIG volatility data. Jarque-Bera is the p-value reported by the Jarque-Bera test for normality.

| Observations | Mean | Std Dev | Skew | Kurt | Jarque-Bera |
|---|---|---|---|---|---|
| 1-1168 | -0.40 | 0.96 | 0.15 | 1.03 | 5.78e-13 |
| 1169-1854 | 0.45 | 1.13 | -0.18 | 0.39 | 0.016 |
| 1855-1944 | -0.64 | 1.21 | 0.31 | -0.07 | 0.47 |
| 1945-2146 | 0.19 | 1.15 | 0.079 | 0.08 | 0.85 |
| 2147-2250 | 1.20 | 1.30 | 0.34 | -0.23 | 0.33 |
| 2250-2414 | 0.58 | 0.89 | 0.792 | 1.87 | 3.58e-10 |
| 2414-2539 | -0.34 | 0.67 | 0.04 | -0.01 | 0.98 |

autocorrelation function. Figure (9.3) presents a spectral estimate, in the top right hand corner of the figure the 95 percent confidence interval is marked which can be used to assess whether there are periodic components present in the data. The ACF and spectral estimate were typical of long memory time series. A visual examination of these diagnostic plots for the 16 series suggested they were all of the long memory type.

### 9.3.2 $d$ Estimates

Both procedures outlined in Section (8.2) were applied to these series. Both procedures begin with the estimation of $d$ and a Beran (1992) goodness-of-fit test of the full series. These two results are presented in Columns "d Est" and "Beran" of Table (9.2) respectively. As reported by other authorities a $d$ estimate close to 0.4 was appropriate for these series. Despite the literature which states that realized volatilities usually present a long memory behavior and that a visual examination of the time series plot, ACF and periodogram suggested all 16 series were of the long memory type, for 11 of the 16 (AA, BA, CAT, GE, HP, IBM, INTC, KO, MSFT, PFE and WMT) the null hypothesis of an FI(d) with $d$ estimated from the data was not accepted at a level of

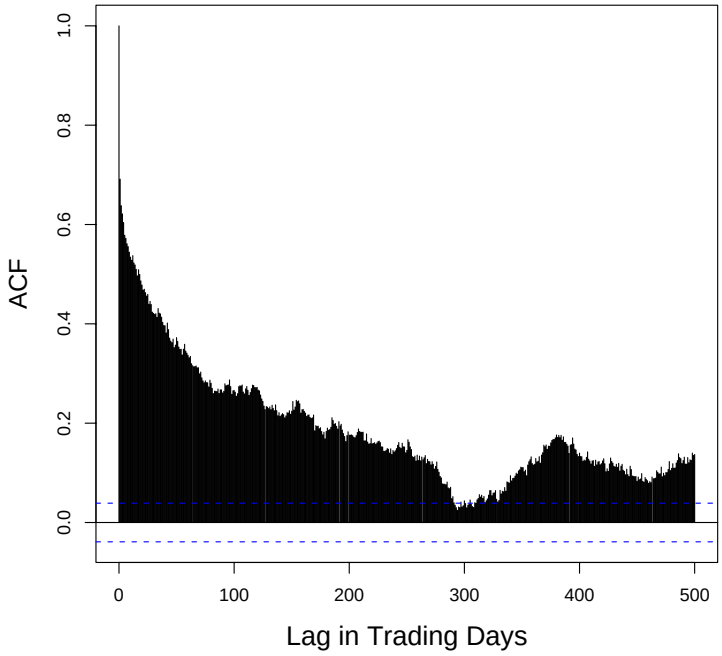


Figure 9.2: ACF of AIG realized volatility series.

5%.

### 9.3.3 Numerical – ART with Beran Test

Next, we applied ART to the full series. Figure (9.1) presents the breakpoints in graphical form superimposed on the plot of the data. Column "Period" of Table (9.3) presents the range of trading days which form each regime. The first line in the table is the full series, the subsequent lines are the regimes.

We then estimated $d$ and performed two Beran tests for each regime. For AIG column "d Est" of Table (9.3) reports the $d$ estimates, column "d=0.40" reports the goodness-of-fit using the series $d$ estimate, column "d=d(t)" reports the goodness-of-fit using the within regime $d$ estimates.

There are two ways to interpret the data in Table (9.3). The first is to examine the

Table 9.2: For the 16 stocks in the sample, column "d Est" reports $d$ as estimated by the Haslett and Raftery (1989) estimator. Column "Beran" reports the p-value for the Beran goodness-of-fit test when applied to the full series. Column "Const $d$" reports the p-value obtained when the Beran (1992) test was used to test if $d$ within the regimes reported by ART were different from $d$ for the full series. Column "Imp $p$" reports the p-value obtained when the number of times the Beran test showed an improved fit by using the regime $d$ estimate instead of the series $d$ estimate was used. Column "ORT" is the test statistic from the method of Ohanissian et al. (2008) a single asterisk ($*$) marks the result which was significant at the 0.05 level. Column "Graphical" reports the p-value for the graphical procedure for testing for a constant $d$ presented in this thesis.

| Stock | d Est | Beran | Const $d$ | Imp $p$ | ORT | Graphical |
|---|---|---|---|---|---|---|
| AA | 0.42 | $2.65 \times 10^{-6}$ | 0.006 | 0.03 | 2.68 | $1.54 \times 10^{-5}$ |
| AIG | 0.40 | 0.14 | 0.009 | 0.04 | 4.02 | 0.068 |
| BA | 0.40 | 0.02 | 0.10 | 0.18 | 1.63 | 0.0001 |
| CAT | 0.41 | 0.002 | 0.0007 | 0.15 | 5.32 | $6.02 \times 10^{-6}$ |
| GE | 0.44 | 0.04 | 0.008 | 0.11 | 4.98 | $3.32 \times 10^{-5}$ |
| GM | 0.36 | 0.09 | 0.34 | 0.03 | 5.20 | 0.0004 |
| HP | 0.44 | 0.002 | 0.04 | 0.04 | 0.50 | 0.004 |
| IBM | 0.44 | $8.93 \times 10^{-6}$ | 0.02 | 0.05 | 1.13 | 0.02 |
| INTC | 0.46 | $3.59 \times 10^{-7}$ | 0.03 | 0.30 | 2.20 | 0.03 |
| JNJ | 0.40 | 0.16 | 0.46 | 0.04 | 3.62 | 0.02 |
| KO | 0.42 | $2.57 \times 10^{-6}$ | 0.02 | 0.04 | 4.40 | $3.74 \times 10^{-10}$ |
| MRK | 0.39 | 0.34 | 0.06 | 0.03 | 3.89 | 0.006 |
| MSFT | 0.46 | 0.008 | 0.43 | 0.007 | $11.24^{*}$ | 0.02 |
| PFE | 0.42 | $4.43 \times 10^{-6}$ | 0.43 | 0.007 | 3.88 | 0.0001 |
| WMT | 0.42 | $2.95 \times 10^{-12}$ | 0.02 | 0.007 | 4.15 | 0.001 |
| XON | 0.44 | 0.09 | 0.04 | 0.04 | 1.26 | 0.004 |

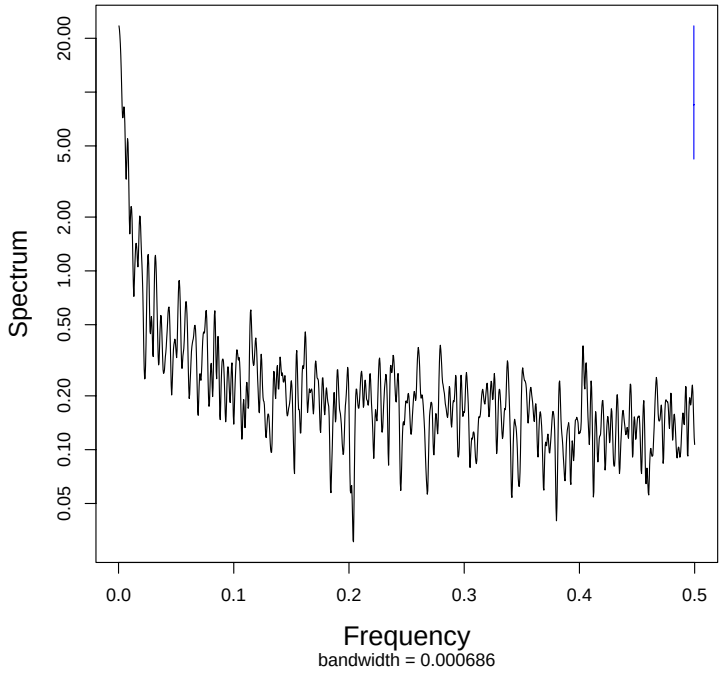


Figure 9.3: Spectral estimate obtained using modified (5,5) Daniell smoothers. The short line in the top right hand corner of the graph is the 95 percent confidence interval.

regimes for which the Beran test fails to accept the null hypothesis when the value of $d$ was held at the series value (column "$d = 0.40$" of Table 9.3). We then check whether the null hypothesis was not rejected if $d$ was set to the value of $d$ as estimated for that regime.

In the case of AIG, the series could be adequately modeled by an FI(0.4) process as the Beran test reported a p-value of 0.14, which indicated the null should not be rejected. When we applied this method of evaluating if $d$ is constant to the data, the first regime reported by ART was clearly anomalous. The Beran test reported a p-value of 0.003, indicating a clear non-acceptance of the null, if the regime was tested against the series $d$ estimate. However, when the first regime was tested against the within regime $d$ estimate ($d = 0.25$) the Beran p-value was 0.54. This clearly indicated an

FI(0.25) model provided a better fit for the first regime.

Thus, for AIG, ART had reported a structural break in the data at the first breakpoint.

If we count the number of regimes for which the null of the regime being an FI(d), with $d$ as estimated for the full series, was not accepted at either the five or one percent levels and use a simple binomial distribution, we can assign a conservative p-value to the null hypothesis of the full series being an FI(d) with a single value of $d$. These p-values are listed in column "Const $d$" of Table (9.2). As can be seen, for the constant $d$ approach, the null was not accepted for 11 of the 16 stocks.

Alternatively, we could use the data from Table (8.3) and estimate the probability that the number of regimes which showed an improved fit were due to chance. Table (8.3) represents the empirically determined distribution of improvement in fit as reported by the Beran test when using the within regime $d$ estimate rather than the full series $d$ estimate for 1000 simulated FGNs with $H = 0.9$ which are equivalent to FI(0.4) series. In the case of AIG an improved fit was reported for all seven regimes.

A simple approach to assess whether the changes in p-values for AIG were the result of a better fit or simply random events was to assume that changes in p-value were distributed binomially. We can then calculate the probability of seven improvements being a random event.

The $p$-values for this variation of the Numerical method are presented in column "Imp $p$" of Table (9.2) for all 16 stocks. As can be seen, for the improved $p$-value approach the null hypothesis was not accepted for 12 of the 16 stocks.

### 9.3.4 Graphical – Bivariate Distribution of $d$ with Regime Length

The Graphical method differs from the Numerical method in that it establishes through simulation the empirical bivariate distribution of regime length and $d$ estimate for FI(d) series. As noted earlier structural break methodologies tend to report breaks in FI(d) series where no breaks exist. From Table (8.1) we can see that for simulated 2500 data point FI(d) series with $d = 0.40$, which is typical for financial data, on average, ART

Table 9.3: Column "Period" is the period under test in trading days from the beginning of the sample. Column "$d$ Est" reports the $d$ estimates for the full series and the regimes reported by ART. Column "$d = 0.40$" reports the p-value reported by the Beran (1992) test for $d$ as estimated for the series. Column "$d = d(t)$" reports the p-value of the Beran (1992) test for the regimes of AIG realized volatilities.

| Period | $d$ Est | $d = 0.40$ | $d = d(t)$ |
|---|---|---|---|
| 1-2539 | 0.40 | 0.14 | - |
| 1-1168 | 0.25 | 0.003 | 0.54 |
| 1169-1854 | 0.36 | 0.35 | 0.46 |
| 1855-1944 | 0.30 | 0.62 | 0.73 |
| 1945-2146 | 0.37 | 0.76 | 0.79 |
| 2147-2250 | 0.35 | 0.38 | 0.39 |
| 2251-2414 | 0.33 | 0.43 | 0.62 |
| 2415-2539 | 0.27 | 0.81 | 0.85 |

reported about seven breaks. All of these breaks were spurious as the data generating process was uniform throughout the series.

The results for AIG are presented in Figure (9.4). As we are interested in whether $d$ has changed, the figure presents the empirical conditional bivariate distribution of $d$ given regime length when the null hypothesis of FI(0.40) was known to be true. This is somewhat different to the form these distributions were presented in Chapter (8), but should be easily understandable. The solid and three sets of dashed lines present the empirically determined median, 75%, 95% and 99% confidence intervals. The "A" symbols represent the seven AIG data points. This method automatically took account of the problem that the Haslett-Raftery estimator was biased in short series, which was seen in Figure (9.4). Visual inspection of this graph showed the 1168 trading day regime (right most "A" in Figure 9.4) had a statistically significant different $d$ value to the full series. Again, we can use a simple binomial distribution to obtain a conservative p-value of whether the null was likely to be true.

The probability that we have at least one observation in seven outside the 99 percent confidence interval is

$$\sum_{i=1}^{7} \binom{7}{i} (0.99)^{7-i} (0.01)^{i}$$

which in this case is 0.068.

Column "Graphical" of Table (9.2) presents the calculated p-values for this method for all 16 stocks. As can be seen, the null hypothesis was not accepted for 15 of the 16 stocks. The null hypothesis was not rejected for AIG. The calculation of the p-value took no account of how far the points were outside the empirical confidence intervals. A visual inspection of Figure (9.4) suggests the null hypothesis would not be accepted for AIG if a more sophisticated calculation was carried out.

### 9.3.5 Comparisons with Other Tests and Procedures

For comparison purposes we present two alternative views of the data; the new test of Ohanissian et al. (2008) and rolling window estimates of $d$ such as those used by Cajueiro and Tabak (2005).

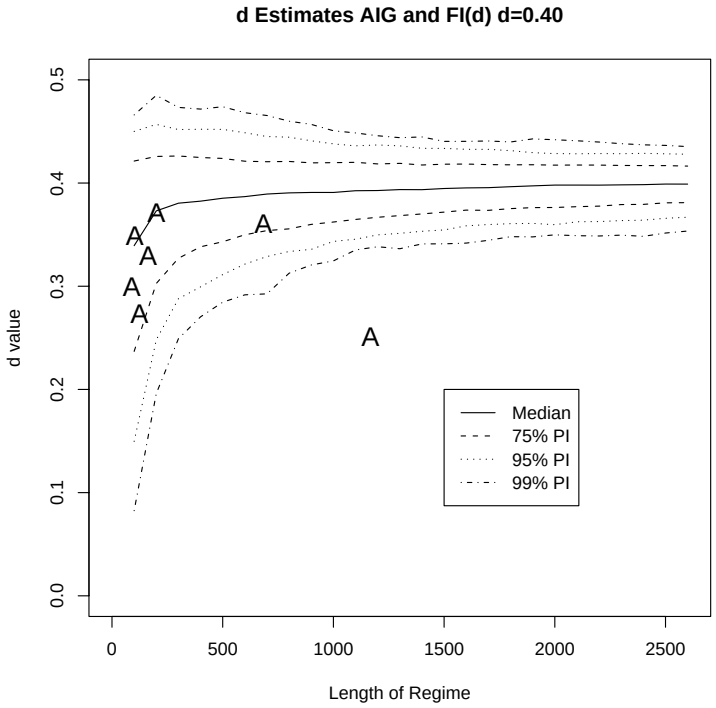


Figure 9.4: Conditional bivariate distribution of $d$ given regime length. The AIG regimes are represented with the symbol "A".

The results for the new test of Ohanissian et al. (2008) are presented in column "ORT" of Table (9.2). The Ohanissian et al. (2008) test only rejected the null of a FI(d) for the MSFT stock. We believe the reason for the lack of rejections to be a consequence of their choice of the GPH estimator. Their test appears soundly based in theory. In practice, for series of the length we are dealing with either the Haslett-Raftery or Whittle estimators would be more suitable.

It is not possible to obtain an estimate for $d$ at a point. The best that can be done is to estimate $d$ in some data window about the point for which we want to know the value of $d$. Panels (a) and (b) in Figure (9.5) present the results of the rolling $d$ estimates using two window sizes as reported by ART for the AIG data. The 1168 trading day window reported in panel (a) was the length of the first regime. The evidence from this

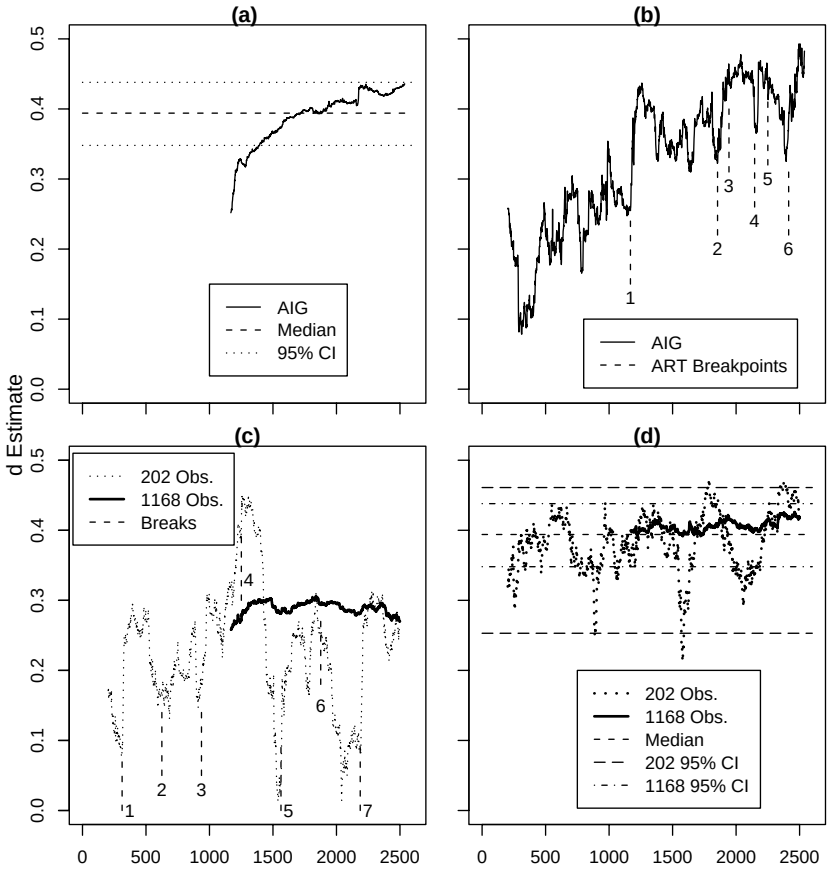


Figure 9.5: Panel (a) presents the rolling $d$ estimates for window length 1168 trading days and panel (b) 202 trading days. Horizontal lines in panel (a) are the median and 95% confidence intervals for FI(0.4) series. The vertical dashed lines in panel (b) are locations of the six breakpoints reported by ART. Panel (c) presents the rolling estimates of $d$ for the same two window lengths for a seven break multifractal (details in text) with breakpoints marked by dashed vertical lines. Panel (d) a siumlated FI(0.40) series with two window lengths with median and 95% confidence intervals.

rolling window supports the results of our two new procedures which reported that $d$ was not constant. The lowest $d$ estimate reported for any 1168 trading day period was the first regime reported by ART.

In panel (b) a 202 trading day window was used. This was the length of the third regime reported by ART and had the highest estimated within-regime $d$ value. The locations of the ART breakpoints are marked and labeled one through six. It is curious that breakpoints 1, 2, 4, and 6 are at sharp changes in the $d$ estimate. One should not over interpret this as there are other sharp changes for which ART did not report a break and two breakpoints (3 and 5) which did not correspond to a sharp change in $d$.

Panel (c) of Figure (9.5) presents the $d$ estimates for the same two data window sizes (202 and 1168 data points) for a simulated multifractal with mean shifts which has eight regimes (seven breaks) alternating between 312 data points at $d$=0.15 and 313 data points at $d$=0.3 and with mean shifts of one standard deviation in terms of the input noise series. In this particular series the 202 data point window was shorter than the regimes and hence at most the window included data from two regimes. The longer data window always covered parts of at least four regimes and did not reflect the series' multifractal structure. The locations of the breaks are marked by dashed vertical lines and labeled 1 through 7. Panel (d) of Figure (9.5) presents the results for a simulated FI(d) series with $d$=0.40 and the same two window sizes.

### 9.3.6 Higher Moments

In Section (8.2) we stated that the "Graphical" method could be applied to higher moments. Figure (9.6) presents the results of the estimates of the standard deviations for AIG. Three of the AIG data points are below the 95% empirical confidence interval, two of which are well below the 99% empirical confidence interval. As the standard deviation is a measure of the variability of the data, for the three regimes with low standard deviation the data were more homogeneous than expected for an FI(d) series.

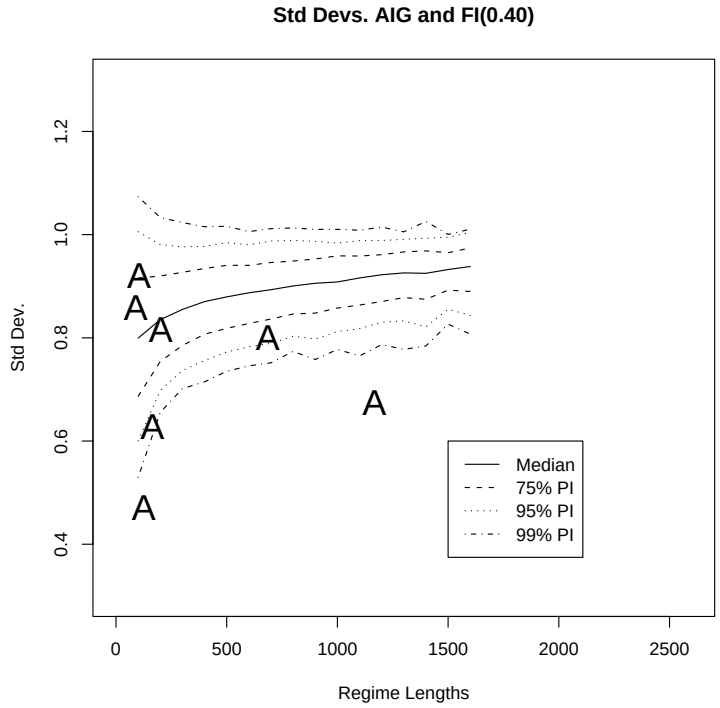


Figure 9.6: Conditional bivariate distributions of standard deviation given regime length for simulated FI(0.40) series. The "A" symbols are the standard deviations of the AIG realized volatility series regimes as reported by ART.

### 9.3.7 CUSUM-Break Distribution

We plotted the CUSUM range against the number of reported breaks for 1,000 simulated series and for the stock. In the CUSUM test the residuals are standardized by dividing by the estimated series standard deviation and the cumulative summation of the residuals is plotted against time. Under the null hypothesis of no structural breaks in the mean and in the absense of serial correlation, the cumulative summation forms a Brownian Bridge usually referred to as the empirical fluctuation process (EFP). The range of the CUSUM test is simply the difference between the maximum and minimum values of the EFP.

The results for AIG are presented in Figure (9.7). As can be seen from the figure

the AIG data point was an extreme value in the distribution. Fifteen of the 16 stocks had extreme values in this type of distribution, with WMT being the exception. It is intuitively easier to understand the extreme value of the CUSUM range if the series underwent sudden significant changes in mean rather than the "quasi-trend" and "quasi-cyclic" behaviour of FGN and FI(d) series.

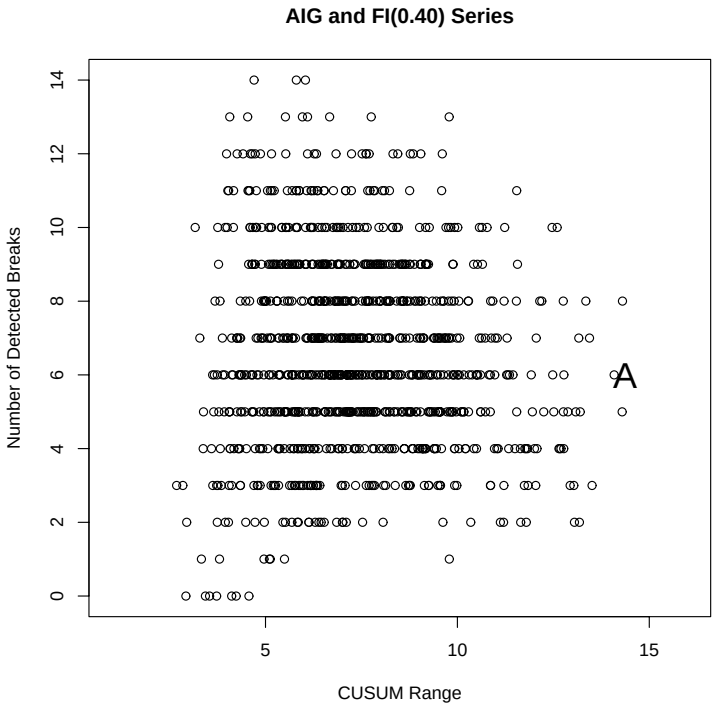


Figure 9.7: Distribution of CUSUM range against reported breaks for simulated FI(0.40) series and the AIG realized volatilities.

### 9.3.8 Historical Events

In their study of these 16 series Scharth and Medeiros (2007) listed six events which they believed influenced stock market volatility. Some of these events were close to instanteous and their effects were felt immediately. Others were events which developed over time and whose effects were felt over a longer period.

Table 9.4: Six major events identified by Scharth and Medeiros (2007) as having an effect on stock market volatility and how many series show a structural break near these dates.

| Event | Approximate Date | No. of Series with Break |
|---|---|---|
| Asian Financial Crisis | 1 Aug 1997 | 4 |
| Russian Financial Crisis | 20 Jul 1998 | 11 |
| Brazilian Financial Crisis | 13 Jan 1999 | 1 |
| NASDAQ Peak | 3 Mar 2000 | 9 |
| Terrorist Attacks | 11 Sep 2001 | 3 |
| DJIA 4-year low | 9 Oct 2002 | 11 |

Table (9.4) presents a list of these events and the number of series for which ART reported a break close to this time. The strongest event was the Russian financial crisis in which 11 of the 16 stocks showed a structural break within 12 trading days of 20 July 1998. Scharth and Medeiros (2007) noted that falling markets tended to have higher volatility than rising markets. Their observation was supported by the fact that nine and 11 series respectively showed a structural break close to the bursting of the technology bubble in early 2000 and the bottoming out of the Dow Jones Industrial Average two and half years later.

Individual stocks also had a number of other breaks. Some of these could be related to known events which one would expect would only affect that stock and perhaps a very small number of related companies. For example, the AIG stock showed a break at 9 April 2001. At this time there were talks between AIG and American General Corporation (AGC). In May 2001, AIG bought AGC for $US 23 billion.

The above observations lend support to the idea that the reported breaks were true structural breaks at which many statistical properties of the regimes changed.

## 9.4 Discussion

In Section (8.1) we noted that when breaks are reported by a structural break location method in FI(d) series that they are fundamentally different from breaks reported in, say, a Markov switching series. Most of this chapter has concentrated on analyzing the long memory properties of the regimes, yet we recall that ART locates candidate breakpoints by detecting shifts in the mean. A change in $d$ and a shift in mean are only related if ART has reported a genuine structural break in the series.

The results presented in Chapter (8) showed that our intuition about the differences between the two types of breaks was indeed correct when our procedures were applied to simulated data. ART usually correctly located the switch points in the two state Markov switching series and the $d$ estimator then correctly reported an estimate of $d$ of zero or close to zero for the regimes. The breaks reported by ART for the FI(d) series were fundamentally different. They were caused by the long excursions away from the mean exhibited by FI(d) series. All of the reported breaks are spurious. Consequently the $d$ estimates in Figure (8.6) reflected the fractional integration of the whole series.

The empirical evidence from Figures (9.1), (9.4), (9.5), and (9.6) and Table (9.3) all point to the first breakpoint being a genuine structural break. Historically this break was at the time of the Russian financial crisis.

As a parameter only has meaning in the context of a model, if $d$ is allowed to be a function of time, as demanded by the evidence, we are led to the so-called multifractal models. However, this leads to three problems. First, a time-varying $d$ nullifies one of the great strengths of the FI(d) models (or FGNs), namely that a single parameter can model the long-range dependence properties of the whole series. On a more technical level, a time-varying $d$ (or $H$) violates the definition of self-similarity. Once a time-varying $d$ is admitted, then any model for the data must of necessity be non-stationary.

Second, as far as we are aware, there are no estimators of $d$ which can handle a time-varying $d$. All assume that $d$ is constant as required by the FI(d) model.

Third, there is the issue of interpretability. It has been difficult to interpret FI(d) series in terms of economic and financial processes generating the volatility data. Con-

sider the implication of an increasing $d$ as observed in the AIG data. An FI(d) model with constant $d$ tells us that any given observation is correlated with all past states of the system into the infinite past. If $d$ is allowed to increase, we have the strange situation where data points separated by longer periods of time were more strongly correlated than those separated by shorter periods of time yet there was no statistically significant cycle present in the data.

Given the arguments of Klemes (1974) it seems reasonable to consider whether estimates of time-varying parameters which were obtained by estimators which assume the parameter is constant, such as those for $d$, have any meaning at all in these series.

The results presented in Figure (9.6) showed that in three cases the data within the regimes were more homogeneous than expected at the five percent level. This was further evidence that the mean shifts reported by ART were, in fact, real and not an artifact of incorrectly analyzed FI(d) series.

While past literature claims FI(d) series model the long memory properties of financial data well, perhaps the most serious drawback to them is on a philosophical level. The breaks in FI(d) series reported by ART are the result of random fluctuations in the data generating process to which no further significance can be assigned. For AIG four of the six reported breaks were correlated with historical events for which a plausible cause and effect mechanism can be proposed.

In previous studies the use of the test of Beran (1992) has often been overlooked. It is a valuable tool in estimating the goodness-of-fit of real data to an FGN or ARFIMA(p,d,q) series. The $d$ estimates presented in the second column of Table (9.2) are comparable with estimates reported by other authorities for similar series. But what has not been previously reported was that often the fit to an FI(d) was poor with the null not being accepted for 11 of the 16 series in our data set. The results from our procedures point to $d$ not being constant coupled with mean shifts as a probable cause of the poor fit. For AIG all seven regimes could be adequately modeled by an FI(d) provided $d$ was allowed to vary with time.

For comparison purposes we have provided two alternative views of the data used

by other authorities.

First, we applied the new test of Ohanissian et al. (2008). These results are presented in the column labeled "ORT" in Table (9.2). We chose four levels of aggregation and the critical values for the $\chi^2$ distribution with three degrees of freedom were 7.82 for $p = 0.05$ and 11.35 for $p = 0.01$. The results showed only MSFT was not a true long memory process on the basis of the Ohanissian et al. (2008) test.

Second, we have used rolling estimates of $d$ (see Figure 9.5). In previous literature there has been no way to determine where one might expect the value of $d$ to change or what length data window should be used. On both questions we have allowed the results of ART to guide us. The two different length data windows showed a statistically significant increase in $d$ with time. The 202 trading day data window showed a distinct break in $d$ estimate at the end of the first regime as reported by ART and smaller changes at breakpoints 2, 4, and 6. These results lend some support to the results of our new methodology. However, the use of rolling estimates of $d$ appears less useful in determining where $d$ may change compared to either of our two procedures.

## 9.5 Conclusions and Future Research

It is now well established that long memory and structural change are easily confused. However, most researchers, particularly in the financial econometrics area, choose to ignore the problem, or simply find it too difficult, empirically, to distinguish between them.

One of the main contributions of this thesis was to propose a new approach and related techniques that can quickly locate potential breaks in the series and test whether the breaks identified are spurious. This is a fundamentally different way to approach the problem of distinguishing between long memory and structural change. It has not previously been proposed and complements the recent approaches proposed by, for example, Ohanissian et al. (2008) and Smith (2005). The particular structural break test method used here, based upon Atheoretical Regression Trees, is fast enough to be practical with the large sized datasets typical in the financial econometrics area. The

new approach was then applied to 16 financial data series to examine whether mean shifts and long memory were linked in realized volatility series. The data were also examined using some existing tests including the new test of Ohanissian et al. (2008).

This chapter set out to examine whether mean shifts and long memory were linked in realized volatility series. In all 16 series we were able to establish that link. However, as Martens et al. (2004) found, even after locating and taking account of the mean shifts the $d$ parameter remained significantly different from zero. We also established that pure FI(d) series did not fit the observed bivariate distributions, a time-varying $d$ was required. However, a time-varying $d$ presents serious unresolved statistical issues as we estimated a time varying parameter with tools which assumed that parameter was constant.

The historical events associated with the ART break dates point to the fact that we have real structural breaks in these data.

The graphical method presented here of determining if $d$ was not constant across the series was shown to be more sensitive to changes in $d$ than using ART in conjunction with the Haslett and Raftery (1989) estimator and the test of Beran (1992). We believe our test could be improved through the use of the Whittle estimator.

Ultimately, however, we must admit that the hypothesis of Klemes (1974), that long memory was a statistical artifact caused by analyzing non-stationary time series with statistical tools that assumed stationarity, has not been fully established. We have established the existence of shifting means in the 16 time series examined in this chapter. However, the resulting model(s) of their data generating processes suggest a multifractal with mean shifts which requires a time varying $d$, and hence creates its own statistical problems in estimating $d$ and logical problems in model interpretation. As a consequence, given that FI(d) series are both stationary and linear, they may well be the "best" approximation currently available for these type of data.

## 9.6 Publications and Presentations

Material from this chapter has been presented at the following conferences and will be submitted in the following paper.

1. W. Rea, E. Mendes, L. Oxley, and M. Reale. Realized Volatility: Long Memory or Shifting Means? An Empirical Study *New Zealand Econometric Study Group Meeting* 4-5 August 2007 Hamilton, New Zealand

2. W. Rea, E. Mendes, L. Oxley, and M. Reale. Are stock market volatility series H-selfsimilar? *Workshop on Non-linear and Complex Systems Analysis.* Brisbane, Australia Sep 27-29, 2007.

3. W. Rea, L. Oxley, M. Reale and E. Mendes; Long memory or shifting means? A new approach and application to realised volatility; *In preparation*

# Chapter 10

# Geophysical Time Series

## 10.1 Introduction

In the previous chapter we examined long memory in financial time series for which there is a large and growing literature in econometrics. The origins of long memory series as a field of research lay elsewhere. Long memory time series were brought to prominence by Hurst (1951) through the study of geophysical time series, in particular through the study of hydrological series. Later, FGNs were introduced into applied statistics by Mandelbrot and van Ness (1968), in part, in an attempt to model the famous Nile River data. Mandelbrot and Wallis (1969) studied a wide range of geophysical series and reported long memory to be pervasive in them. Thus geophysical series are a natural subject area to consider when trying to discriminate between true long memory and non-stationarity leading to spurious long memory. In this chapter we focus on geophysical time series and examine closely the long memory properties of these series.

Tree-ring sequences are often used in paleoclimatic studies both as proxies for past atmospheric conditions and as a basis for temperature reconstructions. These series are quite curious as there are a large number of known cyclic influences on climate, yet when spectral analysis is undertaken statistically significant periodicities are rarely found in them. For example, Thomson (1990) applied the newly developed multitaper

spectral analysis methods using discrete prolate spheroidal sequences to the Campito Mountain data (see Section 10.2 below for details). Thomson reported no statistically significant periodicities in the data longer than 5.5 years. Other authors have reported periodicities in tree ring series, for example Keqian and Butler (1998) and Raspopov et al. (2005), but Thomson pointed out that the widely used five percent significance levels were inappropriate for these data, a better level was $1 - (1/N)$ where $N$ is the number of data points. Often all these reported periodicities are, are a "bump" in the periodogram where one would *a priori* expect to find a periodicity.

Spectral analysis of tree ring series is statistically difficult work as there are several solar cycles; the well-known Schwabe-Wolf cycle of approximately 11 years, the Hale cycle of approximately 22 years, the Gleissberg cycle of about 80-90 years, the Suess or de Vries cycle of around 200 years and the Halstatt cycle of about 2300 years. All of these influence the earth's climate and hence growing conditions for trees. These cycles are not constant in period or amplitude. The "11-year" Schwabe-Wolf cycle ranges between a low of close to eight years to a high of around 14 years. This quasi-cyclic behaviour leads to high leakage even with sophisticated spectral analysis tools and so an indistinct peak appears in the estimated spectrum where genuine periodic phenomena exist.

Keqian and Butler (1998) studied solar cycles using tree ring data and reported the effects of changes in solar activity were smoothed out over time, probably because of the thermal inertia of the earth, and the amount of smoothing depended on altitude. In addition to these solar cycles, there are also quasi-regular changes in the circulation patterns in the earth's atmosphere such as the Arctic and North Atlantic Oscillations, see Shindell et al. (2001). At very long time scales there are systematic changes in the earth's orbital characteristics which are known as the Milankovitch cycles, see Imbrie and Imbrie (1979) for an easy introduction.

Some of the data sets we studied are of particular interest for specific aspects of paleoclimatology. For example, the Elk Lake varve sequence, see Section (10.2) for details. The region where Elk Lake is located is at the intersection of three airmasses;

the cold arctic airmass to the north, the dry Pacific airmass to the west, and the warmer and moister Gulf of Mexico-Atlantic airmass to the south and east. Variations occur on all time scales from seasonal to millennial. It is a climatically sensitive area making it ideal as a proxy for climate change. Also, as no streams flow into the lake the majority of the mineralized materials in the varves are interpreted as being deposited by wind. Thus the varves provide a record of wind conditions in the region.

The Elk Lake data is an example of a cyclostratigraphic time series. In the monograph of Weedon (2003) he pointed out (p84) that almost all cyclostratigraphic time series have a red noise background, which is characteristic of long memory time series, the origin of which is difficult to determine.

We examined a range of millennial scale temperature reconstructions. The particular interest in these series arises because since the end of the last ice age the earth's climate has enjoyed a period of relative stability. As the earth is now in a period of rising global temperatures a number of authors have considered the stochastic properties of univariate time series of both atmospheric and oceanic temperatures from instrumental and proxy records on time scales of a few decades to several millenia in an effort to estimate the natural variability of the earth's climate. This sets a baseline of variation which provides a context in which the observed temperature increases can be studied.

In the literature a number of authors have considered fractionally integrated series as models for temperature and closely related time series. Bloomfield (1992) and Bloomfield and Nychka (1992) considered several time series models including FI(d) series to determine whether the observed global warming in instrumental records could be accounted for by natural fluctuations. Bloomfield and Nychka (1992), in particular, concluded the observed rate of temperature rise could not be accounted for by a stationary FI(d) series. Beran (1994, pp29,30) summarized some studies of long memory in instrumental temperature records. Stephenson et al. (2000) considered FI(d) models among several others for the North Atlantic Oscillation and concluded an ARFIMA(1,0.15,0) model fitted the data best. Baillie and Chung (2002) considered

long memory in several tree ring series which are often used as temperature and precipitation proxies and in climate reconstructions. Baillie and Chung found the series they examined to be very well described by FI(d) series with the exception of the period 1800AD to the present in two of their four data sets. Overland et al. (2006) considered three models of the North Pacific Ocean sea surface temperatures; AR(1), FI(d) and a square wave oscillator. Overland and his co-authors could not establish the statistical primacy of any of the three models. Mills (2007) considered in detail long memory in the Moberg et al. (2005) Northern Hemisphere temperature reconstruction. Mills tentatively suggested the evidence favoured a shifting trends in temperature model over true long memory.

Some of these authors were aware than statistical long memory could be caused by the series being non-stationary. However, as indicated above, distinguishing between a mean-reverting non-stationary series and true long memory was difficult with the statistical tools they had available.

In this chapter examine we examine a number of these series. For reasons of space we can only present one analysis in detail, some further results are presented in Appendix (B).

## 10.2 The Data Sets

A number of data sets were studied. We list the data sets analyzed. Most of these are available through the American National Ocean and Atmosphere Administration's web site.

`http://www.ncdc.noaa.gov/paleo`

A number of these data sets were discarded from the study sometimes after a significant amount of analysis had been undertaken. These are listed first with the reasons they were discarded.

### 10.2.1 Discarded Data Sets

**Baffin Island:** A 1241 year summer temperature reconstruction between 752 and 1992 AD based on laminae thickness of sediments from Donard Lake, Baffin Island, Canada by Moore et al. (2001). The data set was problematic in that it had high skewness and was clearly non-normal data. This could have affected the reliability of the Whittle estimator.

**Burgundy:** A 633 year temperature reconstruction based on grape harvest records in the Burgundy region of France by Chuine et al. (2004). It covered a period from 1370 to 2003 AD. This data set showed no evidence of long memory. The ACF decayed quickly, the AIC selected an AR(4) model as adequate and the periodogram was almost flat with only a small rise in power at low frequencies.

**Moberg:** A 1980 year Northern Hemisphere temperature reconstruction by Moberg et al. (2005) based on composite low and high resolution proxies. It covered the period 1 to 1980 AD. The p-value reported by the Beran (1992) test was zero. It was previously analyzed by Mills (2007) who considered an ARFIMA(2,0.45,2) model. The Beran test also reported a p-value of zero for this ARFIMA model. This data was discarded as we could not find a model where the Beran test reported a p-value above zero. Also, the periodogram was atypical of long memory time series. Mills (2007) only considered the long memory properties of this series in the time domain. His paper is a clear example of why long memory time series must be considered in both the time and frequency domains.

**West Greenland:** A 1230 year reconstruction of temperatures in West Greenland from ice core data by Fisher et al. (1996) which covered the period 753 to 1982 AD. This was not examined in detail as the periodogram was atypical of long memory series.

**Yamal:** A 4,000 year summer temperature reconstruction from larch tree data by Hantemirov and Shiyatov (2002) which covered the period 2067 BC to 1996 AD

from the Yamal Peninsula, Siberia. This was discarded as the authors stated that temperature variability at time scales longer than a few centuries was removed by their reconstruction techniques. This was confirmed by spectral analysis which showed exceptionally low power in the low frequencies.

### 10.2.2 Data Sets Analyzed

**Campito Mountain:** An unbroken set of 5405 annual tree ring widths from bristlecone pines (species *pinus longeava*) on Campito Mountain, California. The trees were at an altitude of over 11,000 feet. The data covered the period 3435 BC to 1969 AD. The rings were measured to the nearest 0.01 mm.

The Campito Mountain data set is regarded as one of the standard examples of the concept of a long memory process. The data set was produced by the Tree Ring Laboratory at the University of Arizona. It is available as the data set `camp` in the `tseries` package within `R`.

**Colorado:** A 2247 year temperature reconstruction based on tree ring data by Salzer and Kipfmueller (2005) which covered the period 250 BC to 1996 AD for the Colorado Plateau region. A reconstruction of precipitation was available in the same data file but was not considered.

**Elk Lake:** The Elk Lake varve data was prepared by Dean (1994) of the United States Geological Survey. It was a series of 10,224 annual varves from Elk Lake in north-eastern Minnesota within the upper Mississippi River Basin. Varves are, essentially, layers of debris which collect in the bottoms of lakes and are part of a larger class of data known as cyclostratigraphic records, see Weedon (2003). The Elk Lake data set was unusually long as the lake was formed after the retreat of glaciers at the end of the last ice age.

In the data there were a number of places where the varve thicknesses were constant. Dean (2006) stated that at the base of the varved section there were several short disturbed zones where the varve couplets could not be measured

but the number could be estimated. Each of the estimated varves was then assigned a constant thickness based on the thickness of the disturbed zone. The accumulation rate was not strictly constant but was close to being constant.

**Northern Hemisphere:** A 1283 year Northern Hemisphere temperature reconstruction which covered the period from 713 to 1995 AD by D'Arrigo et al. (2006). The authors offered two reconstructions, one based on standard methodologies and the other based on "Regional Curve Standardization" (RCS). The authors stated that they believed the RCS better captured the low frequency temperature fluctuations. This belief was supported by a higher $H$ estimate in the RCS reconstruction than the reconstruction which used standard methodologies.

**Scotland:** A data set containing three stalagmite thickness series from a cave in Northwest Scotland by Proctor et al. (2002). The thickness data extends to 3600 years before the present.

**Shihua:** Several data sets were available from the Shihua Cave near Beijing, China. We examined the 2650 year record of stalagmite thicknesses and the associated warm season temperature which was reconstructed from the same stalagmite by Tan et al. (2001). Both data sets cover the period 665 BC to 1985 AD. This data was studied by Rea et al. (2007a,b).

The $H$ estimates and Beran (1992) p-values are presented in Table (B.3).

**Tasmania:** A 3592 year reconstruction of warm season temperatures since 1600 BC by Cook et al. (2000) based on Tasmanian Huon pine tree rings.

**Torneträsk:** A 1993 year reconstruction of temperatures at Torneträsk, Sweden from tree ring data by Briffa et al. (1992) which covered the period from 1 to 1993 AD. The data file appears to have been updated since the publication of Briffa et al. (1992).

**Urals:** A 1077 year summer season temperature reconstruction based on Siberian Larch tree ring data by Briffa et al. (1995) which covered the period from 914 to

1990 AD.

**Western USA:** A 1780 year reconstruction of temperatures based on tree ring data in the Western USA by Mann et al. (1998) which covered the period 200 to 1980 AD. Note: The reference to Mann et al. (1998) for this data set was given in Jones and Mann (2004) but this data set seems too long for the reconstruction discussed by Mann et al. (1998).

We present a detailed analysis of the Campito Mountain data in this chapter as the data are regarded as standard example of long memory time series and are of exceptionally high quality. This data set was often attributed in the literature to the late Valmore LaMarche. The earliest paper of his we could find which discusses the Campito data is LaMarche (1974).

Unless otherwise indicated all $H$ estimates were obtained with the Whittle estimator.

A number of other results can be found in Appendix (B).

## 10.3 Basic Time Series Analysis

A plot of the time series is presented in Figure (10.1). The vertical axis is the tree ring thickness in 0.01mm. The horizontal axis is time, the negative numbers are years BC and the positive numbers are years AD.

There are a number of obvious features in the data. The data appear to fit the description of long-memory series. There are local trends and cycles and long periods of deviation from the mean, yet none persists across the whole series.

There were some quite sudden changes which may be structural breaks. One was at about 2800 BC where there was a clear jump from a period of slow growth to one of high growth (excluding the very short drop to near zero). There was a period of high growth for about two hundred years after 1000 AD between two periods of much lower growth.

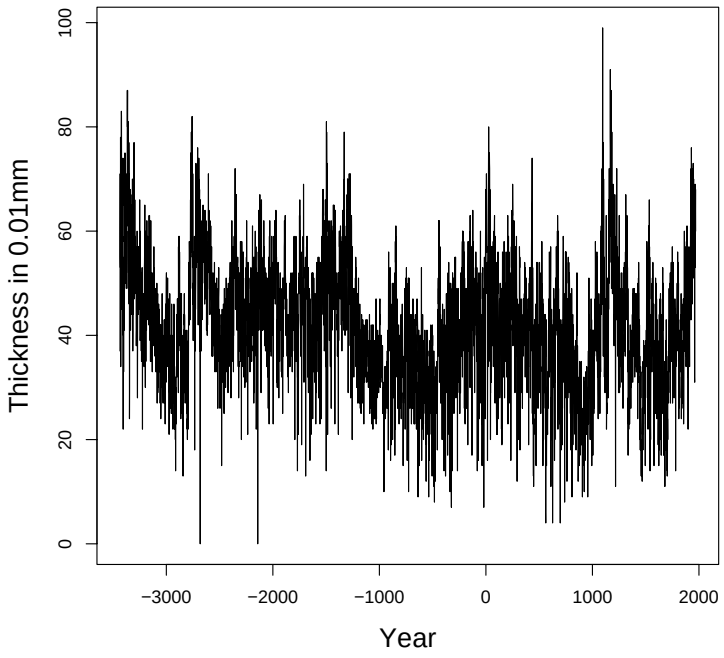


Figure 10.1: Time series plot of the Campito Mountain Data.

There were many downward spikes in the growth rates. These were attributed to short lived atmospheric disturbances caused by volcanic eruptions.

The series gave the impression of heteroskedasticity. There was a period just before 1000 BC where the variability was visually much less than the periods before or after it.

For those searching for evidence of global warming or CO2 fertilization the data show a steady rise towards the end of the series which terminates in 1969.

Some basic statistics on the series are presented in Appendix (B) in Table (B.1) and Figure (B.1). The Jarque-Bera test of normality gave a very clear rejection of the null hypothesis. But the other data in the table, the histogram and the QQ plot all showed that deviations from normality were not particularly large. The skewness was quite small, as was the excess kurtosis. Because of the small deviations from normality,

tests which assume normality (such as the Whittle estimator) should require no special care in interpreting their output.

## 10.4 Fitting an ARIMA Model

In this section we fit a conventional ARIMA model to the data.

Baillie and Chung (2002) and Seater (1993) analyzed the Campito data and reported that for short memory models an ARMA(11,2) model best fitted the data.

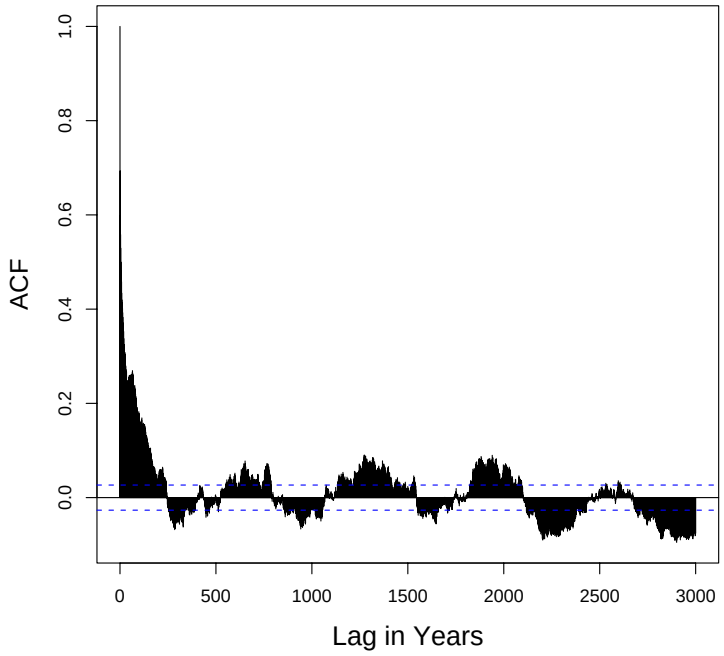


Figure 10.2: The Autocorrelation Function of the Campito Mountain time series plotted to a maximum lag of 3,000 years. The two dashed horizontal lines are the five percent significance levels calculated under white noise assumptions.

The ACF is plotted to 3,000 lags in Figure (10.2). The dotted horizontal lines are the five percent significance levels under white noise assumptions. The PACF (omitted

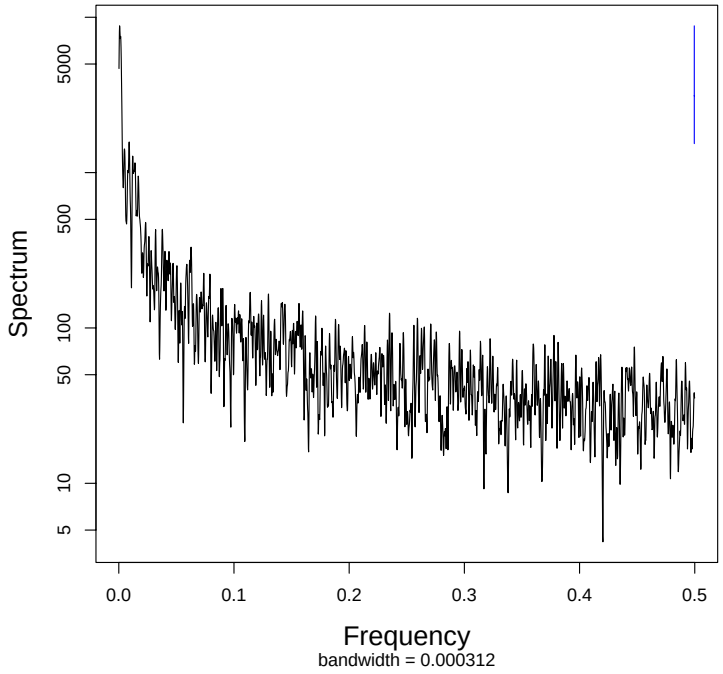


Figure 10.3: The periodogram of the Campito Mountain time series smoothed with (5,5) modified Daniell windows.

for reasons of space) decayed to statistical insignificance after eight lags. These results are not typical of a time series which could be a well described by a short-memory ARMA(p,q) model. The ACF and PACF together show the usual Box-Jenkins program for identifying a good ARMA(p,q) model is of little help.

Taking the ACF plot at face value it showed that the data points 3,000 years apart were statistically significantly correlated. Further, the correlations gave no impression of decaying away towards zero. In fact, the last two periods of negative correlation were stronger than the first three. This differed from simulated FGNs where the ACF often oscillated about the zero line but showed a steady decay with time.

The spectral estimate presented in Figure (10.3) showed no strong cycle in the data. The use of the estimated spectrum for locating periodic components is really

only effective in the case where both the period and amplitude of a cycle are constant. For an introduction to spectral analysis see the texts of either Percival and Walden (1993) or Bloomfield (2000).

The output of `spectrum` with `method="ar"` suggested an AR(14) would fit the data. High order AR models are often required to obtain a good fit when long memory is present. Beran (1994, p147) stated that an FGN process with $H = 0.9$ and 4000 data points would require an AR(21) model to obtain a good fit. Fitting the AR(14) model gave an AIC value of 37847.07. The ARMA(11,2) model favoured by Seater (1993) had an AIC of 37828.43 and so should be favoured.

We conclude that the ARMA(p,q) models can adequately represent the data but at the cost of a large number of parameters.

## 10.5 Long Memory Models

In this section we examine long memory models for this time series.

Baillie and Chung (2002) studied the Campito sequence and determined that an ARFIMA(0,0.45,0) model fitted the data best. The lack of additional short term correlation in the data made it a good candidate for modeling with an FGN.

We applied the 11 estimators studied in Chapter (7) to this series and estimated the goodness-of-fit of the data to an FGN with the 11 estimated $H$ values using the test of Beran (1992). The results are presented in Table (10.1). The Beran test reported an $H$ value close to 0.89 fitted this data best. The maximum p-value was 0.577 for values of $H$ estimated by the aggregated variance and R/S estimators. Nine of the 11 estimators reported $H$ or $d$ values which lay in an acceptable range on the basis of the Beran (1992) test assuming we set our level of statistical significance at 0.05 to reject the null hypothesis of an FGN. The remaining two values could not be tested as the Beran test did not accept $H \geq 1.0$.

Given the results of our simulated FGNs from the study of these estimators in Chapter (7) there were some unexpected $H$ estimates for the Campito data. On the basis of the simulations we expected the aggregated variance, absolute value, boxed

Table 10.1: Column "$H$ Est" presents the $H$ estimates for the Campito Mountain data for each of the 10 estimators of $H$ and the Haslett-Raftery estimator of $d$ converted to $H$ equivalent. Column "Beran p-value" presents the p-values reported by the Beran (1992) test for each of these estimates. The implementation of the Beran test we used could not test the goodness-of-fit for $H$ values above unity. Column "Expected $H$ column" is the expected value that the estimator would report if the Campito data was an FGN with $H = 0.89$. This was calculated by using the data from the simulations presented in Chapter (7). Column "Empirical p-value" column is estimated empirically from the simulated data as explained in the text.

| Method | $H$ Est | Beran p-value | Expected $H$ | Empirical p-value |
|---|---|---|---|---|
| Absolute Value | 0.862 | 0.435 | 0.831 | 0.70 |
| Aggregated Variance | 0.889 | 0.577 | 0.821 | $0.04^{*}$ |
| Boxed Periodogram | 0.914 | 0.509 | 0.849 | $0.01^{**}$ |
| Differenced Variance | 1.089 | - | 0.925 | $0.01^{**}$ |
| Haslett-Raftery | 0.947 | 0.241 | - | - |
| Higuchi | 0.966 | 0.102 | 0.845 | $< 0.001^{***}$ |
| Peng | 0.936 | 0.344 | 0.875 | $< 0.001^{***}$ |
| Periodogram | 1.007 | - | 0.908 | $< 0.001^{***}$ |
| Rescaled Range | 0.892 | 0.577 | 0.816 | 0.36 |
| Wavelet | 0.927 | 0.421 | 0.889 | 0.25 |
| Whittle | 0.876 | 0.540 | 0.890 | 0.15 |

periodogram, Higuchi, and R/S estimators to report a low estimate for $H$. None of these estimators did so. As the Beran test reported that $H = 0.89$ yielded the best fit we used the median value from the simulations with series length 5400 and $H = 0.90$, adjusted for the difference of 0.01 $H$ units, to estimate the value of $H$ which would be reported by each estimator if the data was from an FGN. This value is reported in Table (10.1) in column "Expected H".

The sixth column reports the empirically determined p-value for the $H$ estimate from the Campito data using the distribution obtained from the simulated data in Chapter (7). We do not report values for the Haslett-Raftery estimator as it estimated $d$ not $H$. It is interesting that six of the ten estimators reported $H$ estimates which were statistically significantly higher than their expected values. Thus it appeared that these six estimators were not robust to the specific departures from an FGN found in the Campito data. This was despite the fact that the Beran (1992) goodness-of-fit test reported a maximum p-value of 0.577 for an $H$ value about 0.89, suggesting a very good fit.

## 10.6 The ART of Campito Mountain

In this section we analyze the breakpoints as reported by ART and make some comparisons with the BP breakpoints.

Figure (10.4) presents the regression tree reported by ART. Table (10.2) presents the break dates for both ART and the BP. Figure (10.5) presents the Campito data with the breakpoints marked by dashed vertical lines; panel (a) presents the breaks as reported by ART while those in panel (b) are for the breaks reported by the BP. Figures (10.6) through (10.8) show the twelve individual breaks reported by ART. Table (B.2) presents some basic statistics for the regimes.

The first thing we should notice is that the ART breakpoints are much more irregularly spaced than those reported by the BP. The `mincut` parameter was left at the default value of five. Thus ART could, in theory, report regimes with as few as five data points. BP had an equivalent parameter simply called `h`. It was the minimum segment

Table 10.2: Comparision of break dates reported by ART and the BP for the Campito Mountain data.

| ART Break Date | BP Break Date | BP Confidence Interval |
|---|---|---|
| -3283.5 | - | - |
| -3099.5 | -3088 | -3115 – -3082 |
| -2791.5 | -2789 | -2793 – -2788 |
| -2573.5 | - | - |
| - | -2411 | -2414 – -2391 |
| - | -2138 | -2143 – -2116 |
| -1543.5 | -1544 | -1550 – -1518 |
| - | -1261 | -1262 – -1241 |
| -1215.5 | - | - |
| - | -916 | -921 – -912 |
| - | -454 | -457 – -442 |
| -358.5 | - | - |
| - | 97 | 86 – 164 |
| 440.5 | 440 | 427 – 458 |
| 862.5 | 863 | 854 – 864 |
| 1017.5 | - | - |
| - | 1202 | 1201 – 1208 |
| 1332.5 | - | - |
| - | 1666 | 1661 – 1667 |
| 1862.5 | - | - |

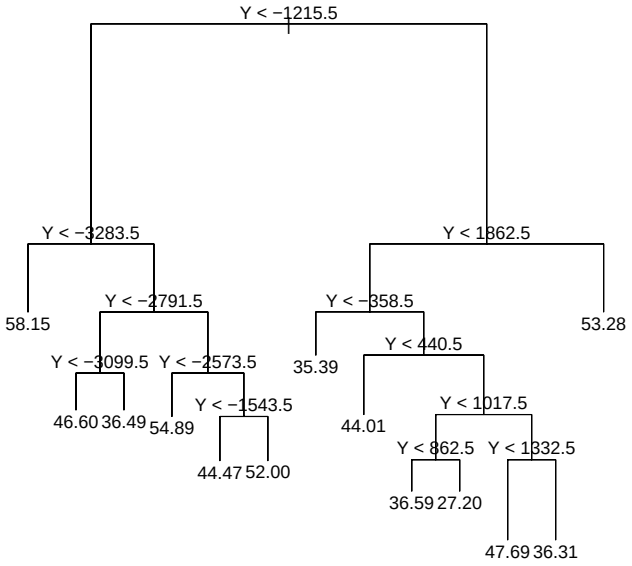


Figure 10.4: Regression Tree for the Campito Mountain Data.

size. Its default was 0.1 of the length of the series, 541 data points in this case. We changed this to 0.05, to allow a minimum regime size of about 270 data points. Thus a number of small regimes reported by ART could not be reported by the BP because the minimum segment size was too high. We lowered $h$ to 0.01 and re-ran the analysis at a cost of over 243 hours of CPU time. This reported 40 breakpoints at the default significance level of 5 percent. This seemed an excessive number but will be referred to as 40BP where appropriate.

We will work backwards through the breakpoints and comment on plausible causes where these are known. These are summarized in Table (10.3).

The most recently reported breakpoint was at 1862.5 AD. 40BP reported a break at 1858 AD. There seem to be two causes for this break which signaled the end of the period known as the Little Ice Age and a return to warmer conditions. Shindell et al.

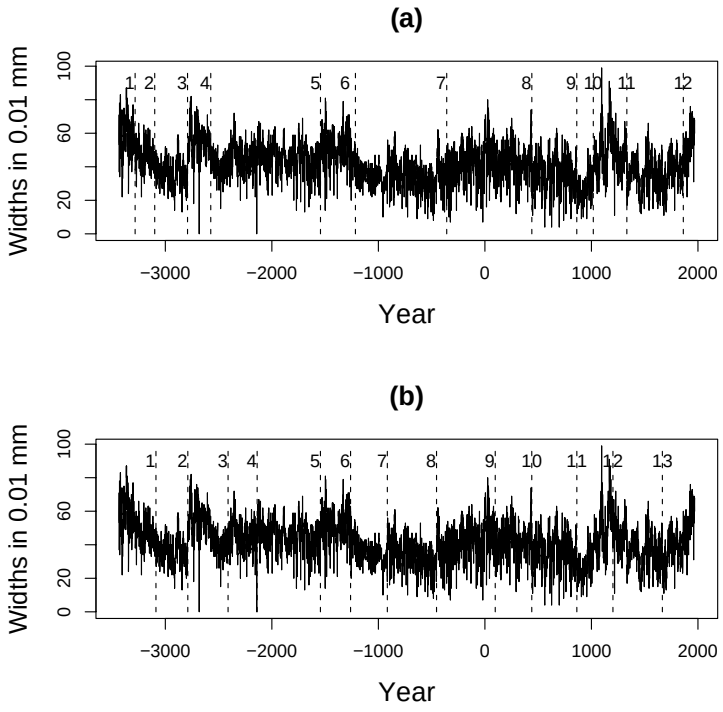


Figure 10.5: Panel (a) Campito Mountain Data with ART breakpoints, panel (b) with BP breakpoints.

(2001) reported there appeared to have been some increase in solar activity after the Maunder minimum. Secondly fossil fuels such as coal, petroleum and natural gas were mined and burned in increasing quantities, steadily increasing the amount of greenhouse gases in the atmosphere and thus trapping heat.

The next most recently reported break was at 1332.5 AD. 40BP reported a break at 1330 AD. The period 1333 to 1862 AD was the period of second lowest average growth rate for the series. Most of this period covered a time known as the "Little Ice Age". Jones and Mann (2004) give the start date of the Little Ice Age to be in the range 1300 to 1450 AD and the end between 1850 and 1900 AD. The breaks reported by ART for the beginning and end of this regime fell within these ranges. Reconstructions of solar activity generally showed lower than average irradiance throughout this period. For

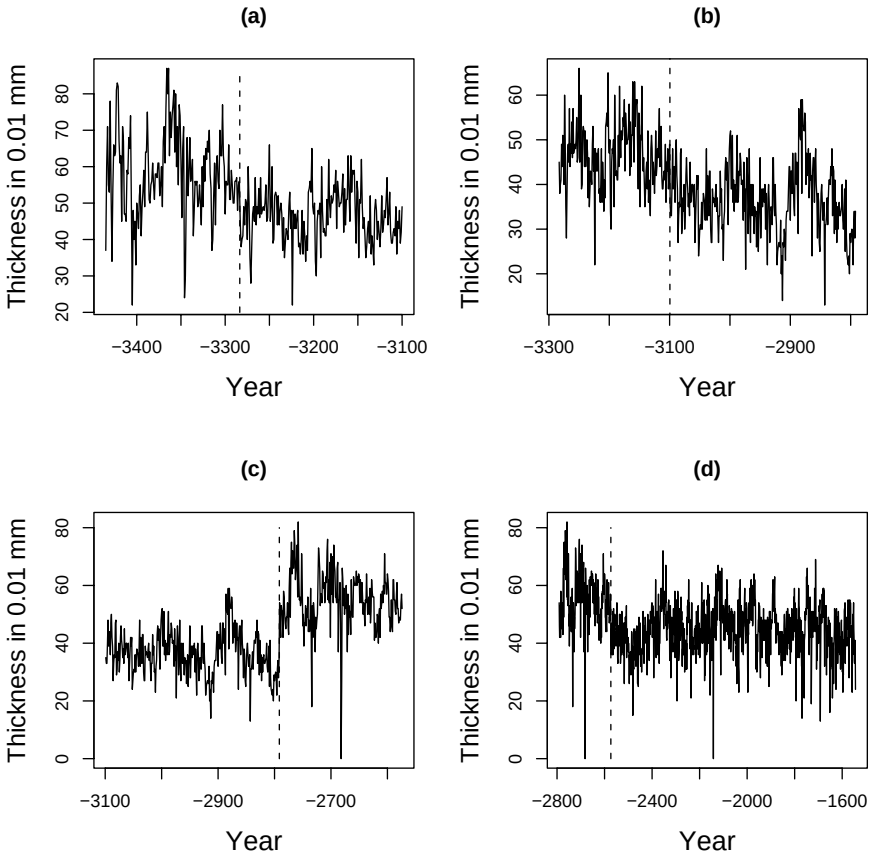


Figure 10.6: Plots of first four breaks in the Campito Mountain series as reported by ART. (a) 3283 BC. (b) 3099 BC. (c) 2791 BC. (d) 2573 BC.

example, see Jones and Mann (2004) Figure (7). A somewhat controversial explanation for part of this cool period is due to Ruddiman (2003) who proposed that the Plague which afflicted Europe in 1347-1352AD caused widespread reversion of cleared farmland to forest and thus reduced the level of greenhouse gases in the atmosphere. This was prolonged by the introduction of new diseases from Europe into the Americas after 1492 which resulted in widespread death and further reversion of croplands to forest.

The next reported break was at 1017.5 AD. The period 1018 to 1332 AD corresponds approximately to the Medieval Climate Optimum. The reconstruction of Jones and

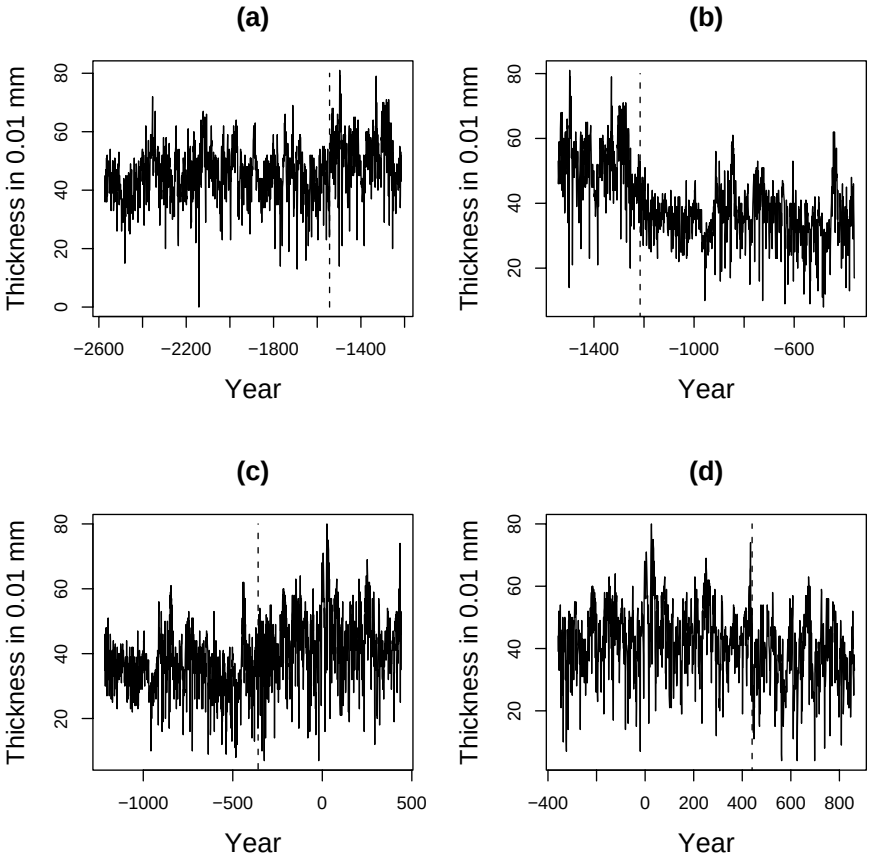


Figure 10.7: Plots of breaks five through eight in the Campito Mountain series as reported by ART. (a) 1543 BC. (b) 1215 BC. (c) 358 BC. (d) 440 AD.

Mann (2004) showed solar irradiance at above average levels for most of this period.

The next reported break was at 862.5 AD. The period 863 to 1017 AD is the period of lowest growth in the series. A possible explanation is drought. Stine (1994) reported evidence of severe prolonged drought in two periods in the Sierra Nevada. They were 912-1112 AD and 1210-1350 AD. These dates were obtained by radiocarbon dating. The first of these approximates this period though rather poorly. Hughes and Funkhouser (2003) discussed the response of bristlecone pines to variations in rainfall. Certainly significantly reduced rainfall of the magnitude reported by Stine (1994) would have

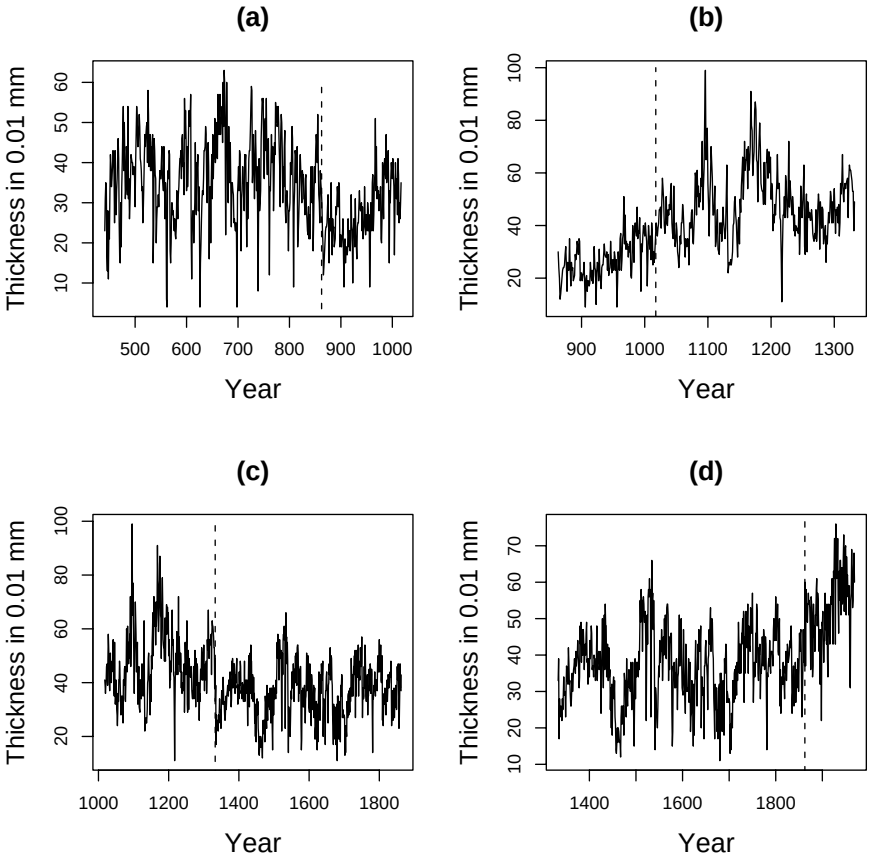


Figure 10.8: Plot of breaks nine through 12 in the Campito Mountain series as reported by ART. (a) 862 AD. (b) 1017 AD. (c) 1332 AD. (d) 1862 AD.

severely curtailed tree growth.

Beyond this it became difficult to find close correlations with other known climate changes. A 2,000 year global temperature reconstruction by Mann and Jones (2003) showed gradually rising temperatures over the period between just before 500 AD and close to 900 AD after having been in decline for the remainder of their reconstruction back to 1 AD. This could correspond to the period 441 to 862 AD. The BP reported an almost identical regime in this period. We must admit there may be circular reasoning at work here as the Campito data has been widely used in climate studies beginning

Table 10.3: The climatic conditions corresponding to the regimes reported by ART for the Campito Mountain tree ring data.

| Period (AD) | Climatic Conditions |
|---|---|
| 1863-1969 | Industrialization and global warming (Jones and Mann, 2004) |
| 1333-1862 | The Little Ice Age (Ruddiman, 2003; Jones and Mann, 2004) |
| 1018-1332 | The Medieval Climate Optimum (Jones and Mann, 2004) |
| 862-1017 | Extreme drought in the Sierra Nevadas (Stine, 1994) |

with LaMarche (1974).

In Figures (10.6) through (10.8) there are some sudden changes in mean ring width. The break at 2791.5 BC, which appears in panel (c) of Figure (10.6), is quite marked. Another jump is seen at 1215.5 BC in panel (b) of Figure (10.7). There is also a sharp drop in panel (c) of Figure (10.8) at the start of the Little Ice Age.

### 10.6.1 ART with CUSUM

The CUSUM plot in Figure (10.9) shows the cumulative sum of the residuals under the null hypothesis that the data are drawn from a single population. A rising line indicates a growth rate above the mean, the falling line indicates a growth rate below the mean. We have superimposed the ART and BP breakpoints on this graph. The CUSUM plot is interpreted subjectively, having regard to the confidence bands marked on it. To our eyes it appeared ART had picked more reasonable breakpoints than the BP, but this may have simply been due to the smaller regime sizes ART could detect.

In these plots there are several places where the breakpoints reported by ART appear more physically reasonable than the breakpoints reported by the BP. The first is the regime between ART breaks three and four, which most closely relate to breaks two and three in the BP. ART reported the break at the end of a period of above average growth while the BP included a short period of below average growth. Similarly, ART breaks nine through 12 separate out the above and below average growth periods. In the corresponding period the BP breaks 11, 12, and 13 report regimes which mix above

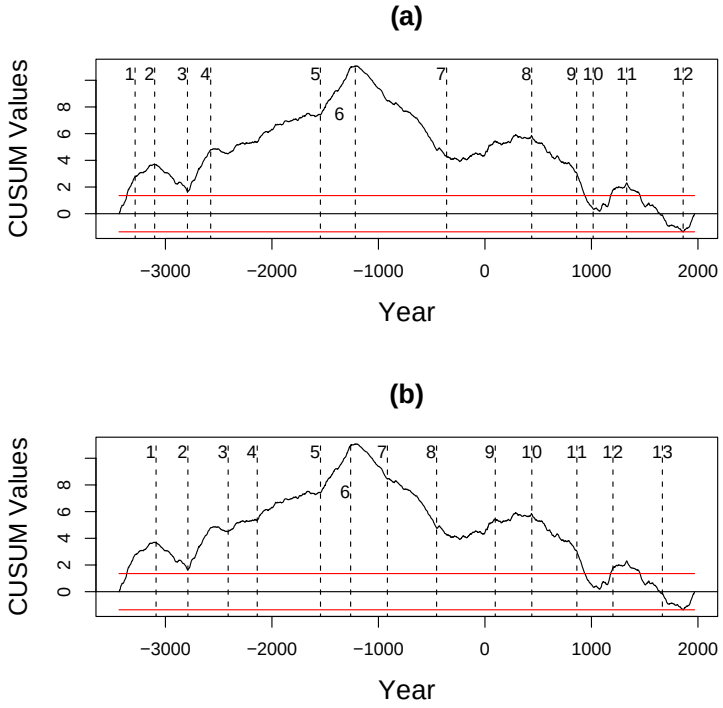


Figure 10.9: Panel (a) CUSUM plot with breakpoints as reported by ART. Panel (b) CUSUM plot with breakpoints as reported by the BP. The lighter horizontal lines about the zero line are the five percent significance levels under null hypothesis.

and below average growth rates.

If we exclude the final regime, which coincides with known anthropogenic changes in the atmospheric greenhouse gas levels, we can see that five of the first six regimes show above mean growth rates while four of the second six regimes show below mean growth. If we were to look for physical reasons for this the most obvious candidate is the shortest of the three Milankovitch cycles, the precession of the earth's axis. The whole period has been one of declining summer insolation in the Northern Hemisphere. See Imbrie and Imbrie (1979) for an introduction to the Milankovitch cycles. This separation of above and below average growth rate periods probably accounts for the strengthening negative serial correlations reported in Figure (10.2).

### 10.6.2 The ART of Regimes

In this section we apply the methods outlined in Chapter (8) and in particular in Section (8.2). Zhang (2007) applied the simple test outlined in Section (8.3) to the Campito data and rejected the null of an FGN with $H = 0.88$ at the five percent level.

In the simulations required for the graphical method all series were 5405 observations in length, with $H$ or $d$ values in the range reported by the various estimators for the Campito data. Usually 10,000 replications were run. The difference is noted if a different number of replications was used. The FI(d) and FGN simulations produced very similar results, only the FGN results are presented. The remainder are omitted for reasons of space.

### 10.6.3 ART Regimes and H Estimates

Table (10.4) presents the results of the Beran (1992) goodness-of-fit test for the Campito regimes using both the within regime $H$ estimate and the full series $H$ estimate. The null hypothesis of an FGN or FI(d) was not accepted at the $p = 0.05$ level for the 1018 – 1332 AD regime and at the $p = 0.01$ level for the 2573 – 1544 BC regime. The longest regime (2573 to 1544 BC) was statistically significantly different from an FGN with $H$ estimated either from the series or the regime. Two other regimes were statistically significantly different from an FGN on one of these measures. As seen from Table (10.4), the period 1018 to 1332 AD would fit an FGN provided the value of $H$ was estimated from the regime rather than from the series.

This result suggested we break the Campito series into three regimes: the anomalous 2573 to 1544 BC data, and the data before and after this period. The results are presented in Table (10.5). At the five percent level the null of long memory was not accepted for the first two periods which collectively represent approximately 35 percent of the data.

Table (10.6) presents estimates of $H$ for the regimes as reported by ART and the cumulative $H$ estimate from the beginning of the series until the end of the current regime. A curious feature of this table is that after the first regime the cumulative $H$

Table 10.4: P-values returned by the Beran (1992) goodness-of-fit test for $H$ variable with time and $H$ constant for the Campito data.

| Period | Regime | $H = 0.876$ | $H = H(t)$ |
|---|---|---|---|
| | $H$ Est | p-value | p-value |
| 3435 to 3284 BC | 0.828 | 0.363 | 0.445 |
| 3283 to 3100 BC | 0.684 | 0.104 | 0.343 |
| 3099 to 2792 BC | 0.810 | 0.520 | 0.429 |
| 2791 to 2574 BC | 0.793 | 0.128 | 0.030* |
| 2573 to 1544 BC | 0.819 | 0.003** | 0.018* |
| 1543 to 1216 BC | 0.796 | 0.186 | 0.469 |
| 1215 to 359 BC | 0.856 | 0.224 | 0.394 |
| 358 BC to 440 AD | 0.916 | 0.416 | 0.439 |
| 441 to 862 AD | 0.856 | 0.925 | 0.934 |
| 863 to 1017 AD | 0.856 | 0.948 | 0.944 |
| 1018 to 1332 AD | 0.966 | 0.013* | 0.329 |
| 1333 to 1862 AD | 0.891 | 0.839 | 0.839 |
| 1863 to 1969 AD | 0.773 | 0.230 | 0.388 |

Table 10.5: P-values reported by the Beran (1992) goodness-of-fit test for three periods in the Campito data divided by the anomalous 2573 to 1544 BC period.

| Period | $H$ | Beran p-value |
|---|---|---|
| 3435 to 2574 BC | 0.868 | 0.047 |
| 2573 to 1544 BC | 0.819 | 0.018 |
| 1543 BC to 1969 AD | 0.883 | 0.641 |

Table 10.6: "Period" gives the absolute dates for each regime reported by ART. "Years" gives the length of each regime in years. "$H$ Est" reports the estimates of $H$ reported by the Whittle estimator. "Cumulative $H$ Est" reports the $H$ estimate for the series from the beginning to the end of the current regime.

| Period | Years | Regime $H$ Est | Cumulative $H$ Est |
|---|---|---|---|
| 3435 to 3284 BC | 152 | 0.828 | 0.828 |
| 3283 to 3100 BC | 184 | 0.684 | 0.834 |
| 3099 to 2792 BC | 308 | 0.810 | 0.885 |
| 2791 to 2574 BC | 218 | 0.793 | 0.868 |
| 2573 to 1544 BC | 1120 | 0.819 | 0.860 |
| 1543 to 1216 BC | 328 | 0.738 | 0.840 |
| 1215 to 359 BC | 857 | 0.796 | 0.855 |
| 358 BC to 440 AD | 799 | 0.856 | 0.849 |
| 441 to 862 AD | 422 | 0.916 | 0.861 |
| 863 to 1017 AD | 155 | 0.856 | 0.867 |
| 1018 to 1332 AD | 315 | 0.966 | 0.872 |
| 1333 to 1862 AD | 530 | 0.891 | 0.879 |
| 1863 to 1969 AD | 107 | 0.773 | 0.876 |

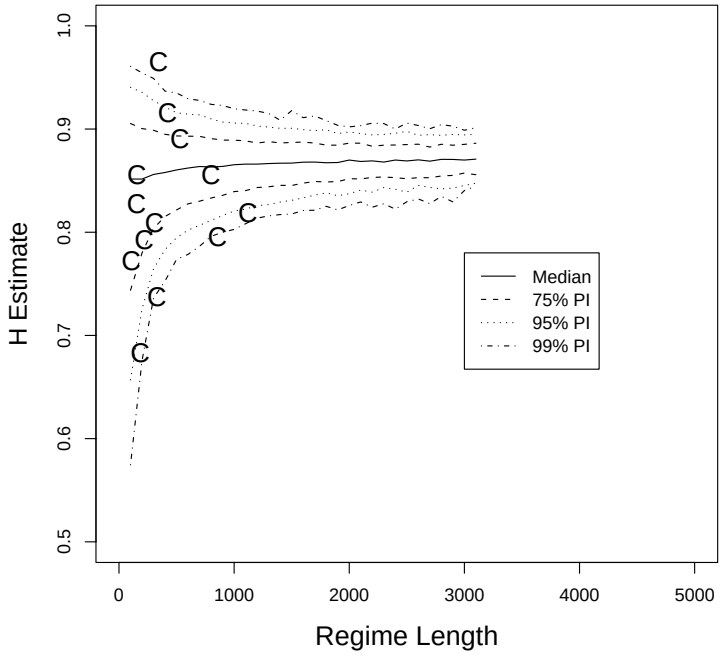


Figure 10.10: The conditional distribution of $H$ estimates given regime length for FGN series with $H = 0.88$. The 13 "C" symbols are the 13 regimes reported by ART for the Campito data.

estimates did not change much even though the within-regime $H$ estimates were quite variable.

Figure (10.10) presents the result of the graphical method presented in Section (8.2) for simulated FGNs with $H$=0.88. Simulations with $H$ values between 0.82 and 0.96 were run but are omitted for reasons of space. As can be seen five of the 13 $H$ estimates for the Campito regimes lie outside the 95 percent confidence interval of the empirically determined conditional bivariate distribution of $H$ given regime length. An anonymous referee for the Journal of Computational and Graphical Statistics pointed out that this only tests a single value of $H$. In one sense this is true. However, the bivariate distribution is dependent on the chosen estimator. The shape of the

distribution changes with simulated $H$ value but rather slowly. If we only change the $H$ value in the simulations by a few hundredths of an $H$ unit to a very good approximation we can regard the shape of the distribution as fixed, it will only be raised or lowered.

As can be seen, the Campito data had within-regime $H$ estimates which were above and below the upper and lower 95 percent confidence intervals. Trying to improve the fit by, say, lowering the tested $H$ value to bring the four lower Campito regime $H$ estimates within the lower 95 percent confidence interval of the distribution would result in at least one more regime $H$ estimate being moved above the upper 95 percent confidence interval. On the other hand, raising the tested $H$ value would result in the four within-regime $H$ estimates which are currently outside the lower 95 confidence interval falling outside the lower 99 percent confidence interval. Thus changing the $H$ value in the simulations can, at best, result in only minor improvements in fit.

On the basis of this set of results it seems unreasonable to claim that a single value of the $H$ parameter characterizes the long memory properties of this series. Further, over one-third of the data does not have a spectral density of the form in Equation (6.16). These results suggest the series is not $H$-self-similar.

### 10.6.4 ART Regimes and Standard Deviations

The standard deviations were estimated in the same set of simulations as in the previous section. These were also standardized to one by dividing by the series standard deviation. Some of the results are presented in Figure (10.11) the remainder are omitted for reasons of space. Five of the 13 Campito data points lay below the empirically determined lower 95 percent confidence interval much the same as we saw in Figure (9.6). Thus within these five regimes the data was more homogeneous than expected if it were generated by an FGN. Also there was one regime within which the data were more variable than expected. The data are also presented in Table (B.2). Only one regime had a higher standard deviation than the series which was about 1.11 times higher than the series standard deviation.

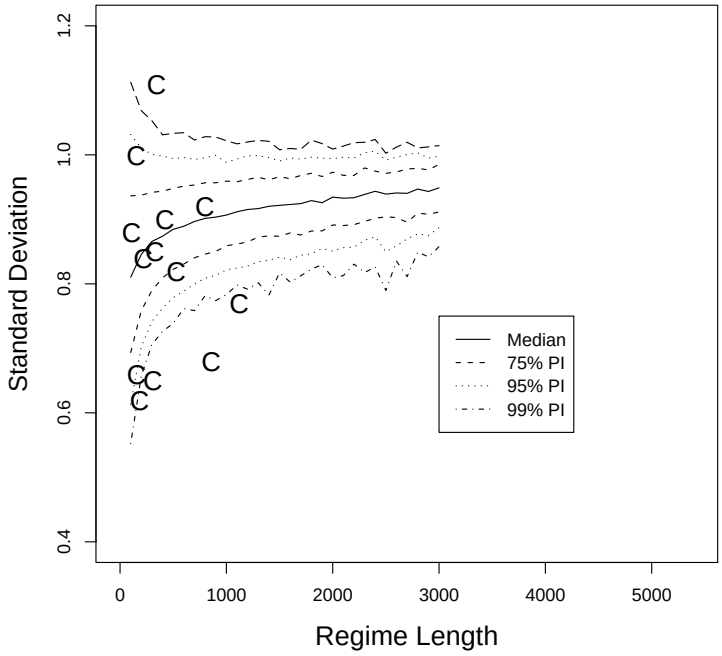


Figure 10.11: The conditional distribution of standard deviations given regime length for simulated FGN series with $H = 0.88$. The standard deviation of the full series has been standardized to one. The "C" symbols are the Campito data points.

### 10.6.5 ART Regimes and CUSUM

In this section we combine the CUSUM test with ART in the same manner as in Section (9.3.7). We ran 1000 replications of simulated FGNs with $H = 0.86$, 0.88, 0.90 and 0.92. We report here the results for the $H = 0.88$ simulations, the remainder are omitted for reasons of space. In Figure (10.12) we have plotted the CUSUM range against the number of breaks reported by both ART and the BP. It is clear that the Campito data are well outside the simulated distribution. The simple test proposed in Section (8.3) can be improved by considering the bivariate distribution of reported number of breaks and CUSUM range rather than reported number of breaks alone.

Thus the Campito data is unusual at at least the $p = 0.001$ level on this measure.

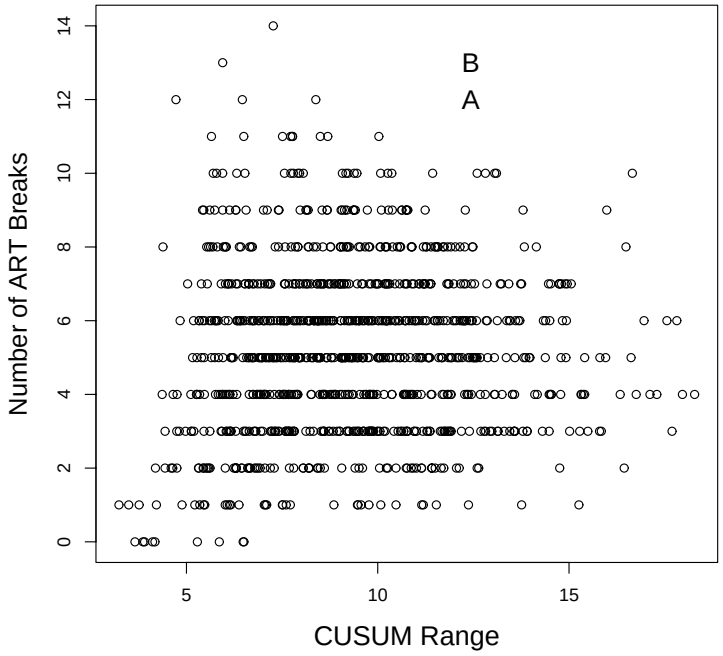


Figure 10.12: Scatter plot of CUSUM range against the number of breaks reported by ART for 1000 simulated FGN series with $H = 0.88$. "A" is the ART data point for the Campito data, "B" the BP data point.

## 10.7 Time Invariant Statistics

In time series analysis one looks for statistics which are invariant over time which can form the basis of a model. In the Campito data we have shown that the key parameter in long memory models, $H$, was not constant with time thus rendering it of little use in modeling and in particular demonstrating to high levels of statistical significance that the DGP is not an FGN. In the analysis of the geophysical series a number of them did have a time invariant measure and that was the fractal dimension.

We recall from the discussion of the $H$ estimators in Section (7.2) that we noted that the Higuchi estimator was not an $H$ estimator, but an estimator of $D$, the fractal dimension. In the simulation study in Section (7.4) we obtained the distribution of $H$

estimates for simulated FGNs which was presented in Figure (7.3).

In Figures (10.13) and (10.14) we present histograms of the estimated $H$ values, as a proxy for $D$, for the Campito data and a simulated FGN respectively. The horizontal axes of these graphs should be examined closely. It is quite startling to see that the real data had a tiny range of $H$ estimates compared to the computer generated data.

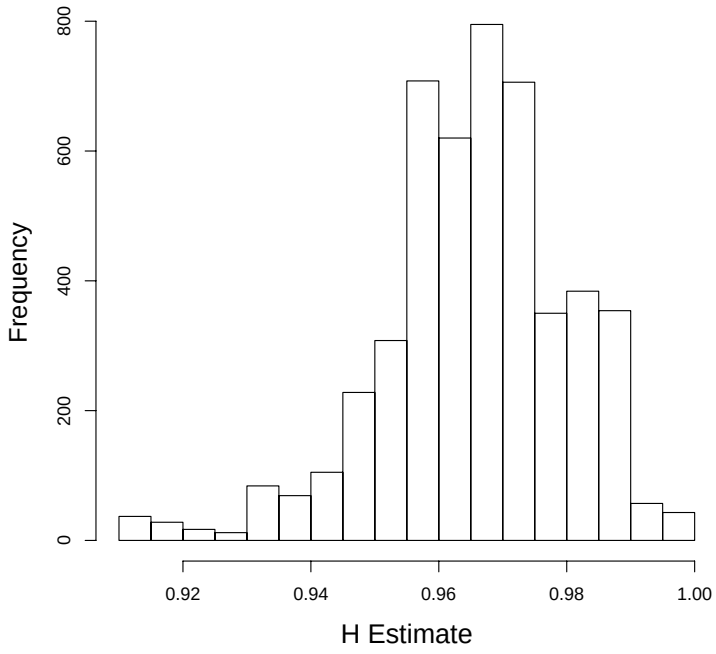


Figure 10.13: Histogram of rolling 500 year $H$ estimates reported by the Higuchi estimator for the Campito Mountain Data.

In addition to the Campito data, the Baffin Island, Colorado Plateau, both Shihua Cave data sets, Tasmania, and the Yamal all exhibited this property. While it is beyond the scope of this thesis to examine this, it is an interesting finding which should be the subject of future research. A suggested title could be "Preservation of the Fractal Dimension" with apologies to Hipel and McLeod (1978).

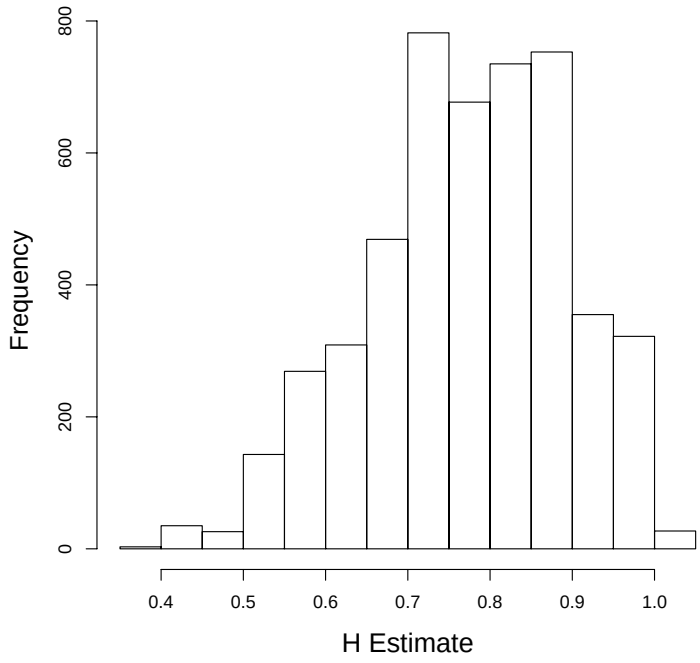


Figure 10.14: Histogram of rolling 500 data point $H$ estimates reported by the Higuchi estimator for FGNs with $H$=0.86.

## 10.8 Discussion

In Section (9.4) we made a number of comments on the results of our investigations of the long memory properties of the realized volatility data which could be repeated here with a few appropriate changes.

There have been long standing doubts that geophysical time series are stationary. There is the highly cited paper of Klemes (1974). But there are others. For example, in a comment on Haslett and Raftery (1989), Glasbey states "Meteorologists have a rule-of-thumb that about 30 years of weather data are optimal to represent current climatic variability because longer periods are affected by drifts in climate."

Of the 12 breaks reported by ART, only three of them (2791 BC, 1215 BC and possibly 1332 AD) resemble anything like a regime shift. The remainder look like

smooth transitions which evolved over a period of time. This interpretation is supported by the CUSUM plot in Figure (10.9) in which most of the transitions between states are rounded and not the sharp "kinks" which are characteristic of an abrupt shift in mean. Thus a regime switching model or some variant with a fixed number of states does not seem appropriate for this data.

The question we ought to consider is – what alternative model can we apply to these types of data? While it is beyond the scope of this thesis to construct such a model we offer some thoughts on possible directions. We have already noted the constant fractal dimension in Section (10.7).

We recall that Thomson (1990) applied a sophisticated spectral analysis tool to the Campito data and reported only very short periodicities in the data. Thomson (1990) also subjected the Campito data to a test of stationarity and concluded it was mildly non-stationary. When the Campito data was split into smaller blocks Thomson (1990) reported "Numerous lines with high apparent statistical significance were found in individual blocks but did not persist between blocks." Thomson explored a number of possibilities for these unexpected results but was unable to establish the statistical primacy of any of them. He did not consider the possibility that periodic phenomena were transient, that is, they were present in some periods but not others. We observed the same phenomena a number of these geophysical time series, not just the Campito data.

At very long time scales the earth's climate system clearly behaves in a non-linear manner, see Gipp (2001). It is also clear that the earth is subject to some highly regular, very long period forcings in the Milankovich cycles and a number of quasi-periodic solar forcings. For many years it was doubted that these tiny changes in the solar forcings could induce the large quasi-cyclic changes observed in paleoclimatic data. For example, Shindell et al. (2001) put the reduction in solar output in the Maunder Minimum at only 0.1 percent. Using global circulation models Bond et al. (2001) established that these small solar forcings could indeed affect the ocean-atmosphere's oscillatory systems.

We propose that the observed lack of statistically significant line spectra in the geophysical series is not because cyclic phenomena do not exist but because they are unstable and do not persist for the whole period of the data. A comparatively easy way to see this is to select a fixed length data window and slide this window across the data estimating the spectrum at each point. Then one can plot the evolution of the spectrum though time. An example is presented in Figure (10.15).

Marked on the spectral estimates in Figure (10.15) are a 7.1 and 23 year period which are statistically significant in some sub-periods but not in others or in the full series. Not presented here are evolutionary spectral analysis plots. When examined in detail it was clear some cyclic phenomena did appear and some lasted for more than a thousand years, such as those marked in Figure (10.15), but none persisted long enough to appear as a statistically significant line in the estimated spectrum for the whole data. Thus some cyclic phenomena were excited at some stages of these periodic forcings but not at others. At long time scales the evolution of the very low frequencies at the back these evolutionary spectral analysis plots are probably attributable to the Halstatt cycle of approximately 2300 years as Thomson (1990) reported the spectrum exhibited an non-periodic evolutionary cycle of 2360 years.

Granger (1966) pointed out that all that was required to produce a red noise spectrum of the type typically observed in economic data was to have a cycle in the data which was longer than the sample. A similar comment could be applied to geophysical data for which we have numerous candidate periodicities at all time scales including ones much longer than the data samples. It is perhaps an unfortunate coincidence that the final spectral estimate resembles that of an FGN when the two processes may have little else in common.

## 10.9 Conclusions

If the goal of the analysis is to find a single statistically parsimonious and mathematically elegant model for all of the data then either fractional Gaussian noise or an ARFIMA model are the models of choice for the Campito series. They are quick to fit,

require few parameters, and are straight forward to use in forecasting. However, the philosophical problem of whether the FGN or ARFIMA models represent some actual physical process or processes can be answered in the negative for this data set.

Regime shift type models, for which ART would be an excellent tool to identify the regimes and breakpoints between them, do not seem an appropriate model either.

Incidental to the main focus of this thesis we proposed that any realistic model of these geophysical time series would need to be based on an understanding of the physics of the various periodic phenomena which appear for a while in the data and then vanish again. Unless one is prepared to accept the limitations of FGNs there are no "off the shelf" time series models which can adequately capture the important features in these series. This is certainly an area which needs more research.

## 10.10 Publications and Presentations

Material from this chapter appeared in Rea et al. (2007a,b) and Zhang (2007). The poster paper on which Rea et al. (2006) was based included the Campito data as a case study.

A presentation was made to the New Zealand Econometric Study Group Meeting 5-6 August (2006) Dunedin, New Zealand: W. Rea, M. Reale, and J. Brown; "Do Long Memory Time Series Have Amnesia?" This presentation won a New Zealand Econometric Study Group - Reserve Bank of New Zealand Award for Econometric Research.

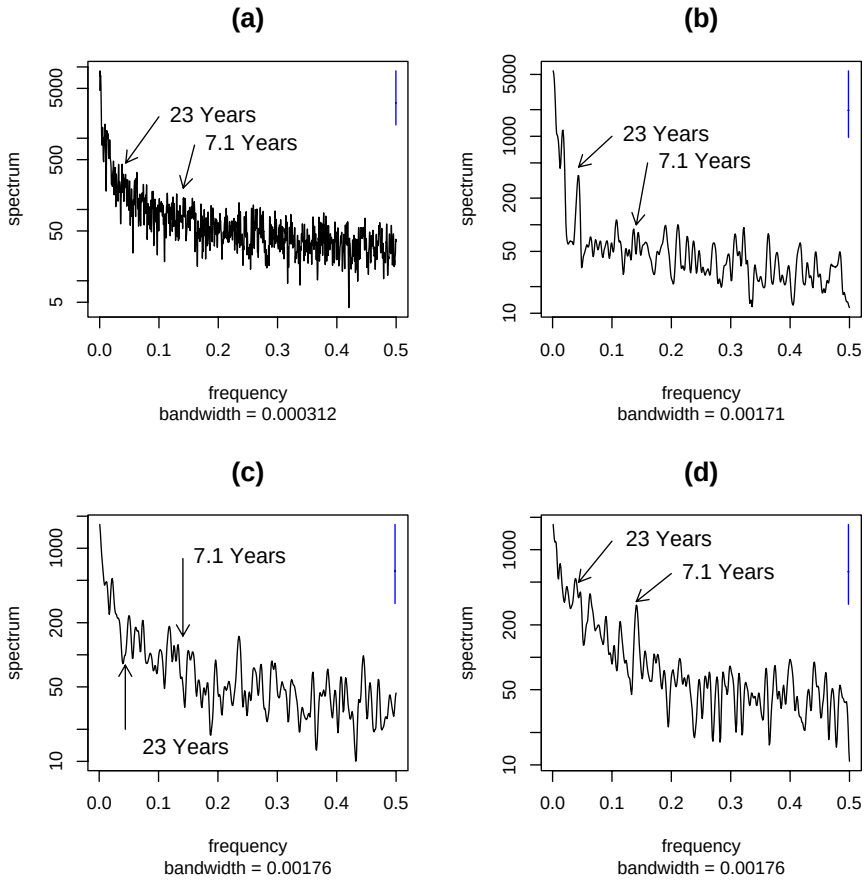


Figure 10.15: Spectral estimates for the Campito Mountain Data. (a) The full series. (b) through (d) 1000 year subsamples beginning at 3135 BC, 2135 BC, and 335 BC respectively.

# Chapter 11

# Summary of Conclusions

## 11.1 ART

This thesis supports the following conclusions about ART as a structural break detection and location methodology.

- In the absence of serial correlations, ART did impose spurious breaks when the series were short but this tendency disappeared as the series became longer. This was seen in single, multiple break and heteroskedastic simulations.
- ART was robust to negative serial correlation and a small amount of positive correlation, but in this regard ART was no worse than the BP.
- Leave-one-out cross-validation can be used for tree selection but was computationally expensive.
- ART reported higher numbers of breaks for series with non-Gaussian noise structures but not excessively so.
- ART's robustness to outliers depended on the outlier location. ART was more robust to outliers away from the ends of the series than those close to the end.
- It is possible to obtain a confidence interval for the breakpoints using the bootstrap. However, because regression trees only fit piece-wise constant functions to

the data, this is currently limited to series without serial correlations and may thus be of little practical use.

- For routine tree selection the BIC was found to be a good choice.
- For series with less than approximately 500 data points leave-one-out cross-validation yielded a worthwhile reduction in spurious breaks reported.
- ART showed greater robustness to a non-normal noise structure than did the BP.
- Plotting the tree gives a visual representation of the breakpoints and the mean values of the regimes reported.
- ART was computationally faster than both CUSUM and the BP. This makes it especially suited to the analysis of long time series.
- We found one weakness in the standard tree growing and pruning method. We have proposed a computationally inexpensive modification to the final tree selection method, which we call "Enhanced Temporal Pruning" which partially overcomes this problem.

## 11.2 Long Memory

### 11.2.1 Estimators

- For series with less than 4,000 data points only the Whittle and Haslett-Raftery estimators should be considered.
- For longer series the Peng, periodogram and wavelet estimators should also be considered.
- The Higuchi estimator is useful if the researcher wishes to recover the fractal dimension of the time series.
- A Beran test should always be run after estimating $H$ or $d$ for a long memory time series.

### 11.2.2 Model Discrimination

1. We have proposed new methodologies to distinguish between true long memory and shifting means which are based on ART. These are fundamentally different from existing tests and procedures.

2. These new procedures worked well in simulated data.

3. We examined a number of financial and geophysical data sets and obtained evidence of non-stationarity in all of them.

4. We showed that a single value of $H$ or $d$ was not adequate to describe the long memory properties of either the financial or geophysical series.

5. In some of the geophysical series we showed that the fractal dimension of the time series was time invariant.

# Appendix A

# Extra Financial Data Results

## A.1 ART with Beran Tests

In this section the results from the ART with Beran test for the remaining 15 stocks are presented.

Table A.1: Estimated $d$ values for the regimes together with the p-value reported by the Beran (1992) test for $d$ as estimated for both the series and the regimes of the Alcoa stock.

| Period | d | d=0.42 | d=d(t) |
|---|---|---|---|
| 1-2539 | 0.42 | $2.65 \times 10^{-6}$ | - |
| 1-310 | 0.20 | 0.04 | 0.54 |
| 311-592 | 0.27 | 0.32 | 0.66 |
| 593-1177 | 0.27 | 0.35 | 0.83 |
| 1178-1555 | 0.42 | 0.53 | 0.53 |
| 1556-1647 | 0.34 | 0.02 | 0.15 |
| 1648-1806 | 0.36 | 0.005 | 0.02 |
| 1807-2351 | 0.34 | 0.07 | 0.18 |
| 2352-2539 | 0.12 | 0.07 | 0.67 |

Table A.2: Estimated $d$ values for the regimes together with the p-value reported by the Beran (1992) test for $d$ as estimated for both the series and the regimes of the Boeing stock.

| Period | d | d=0.40 | d=d(t) |
|---|---|---|---|
| 1-2539 | 0.40 | 0.02 | - |
| 1-678 | 0.26 | 0.28 | 0.81 |
| 679-730 | 0.07 | 0.85 | 0.96 |
| 731-913 | 0.15 | 0.21 | 0.75 |
| 914-1169 | 0.32 | 0.51 | 0.74 |
| 1170-1246 | 0.29 | 0.28 | 0.55 |
| 1247-1542 | 0.37 | 0.06 | 0.08 |
| 1543-1805 | 0.35 | 0.80 | 0.85 |
| 1806-2154 | 0.44 | 0.55 | 0.42 |
| 2155-2262 | 0.41 | 0.58 | 0.57 |
| 2263-2437 | 0.22 | 0.02 | 0.33 |
| 2438-2539 | 0.22 | 0.07 | 0.41 |

Table A.3: Estimated $d$ values for the regimes together with the p-value reported by the Beran (1992) test for $d$ as estimated for both the series and the regimes of the Caterpillar stock.

| Period | d | d=0.41 | d=d(t) |
|---|---|---|---|
| 1-2539 | 0.41 | 0.002 | - |
| 1-72 | 0.13 | 0.001 | 0.26 |
| 73-915 | 0.27 | 0.0001 | 0.06 |
| 916-1178 | 0.25 | 0.18 | 0.73 |
| 1179-1245 | 0.35 | 0.75 | 0.75 |
| 1246-1536 | 0.28 | 0.61 | 0.91 |
| 1537-1804 | 0.35 | 0.66 | 0.75 |
| 1805-2349 | 0.44 | 0.66 | 0.59 |
| 2350-2539 | 0.24 | 0.75 | 0.93 |

Table A.4: Estimated $d$ values for the regimes together with the p-value reported by the Beran (1992) test for $d$ as estimated for both the series and the regimes of the General Electric stock.

| Period | d | d=0.44 | d=d(t) |
|---|---|---|---|
| 1-2539 | 0.44 | 0.04 | - |
| 1-81 | 0.11 | 0.63 | 0.82 |
| 82-261 | 0.29 | 0.75 | 0.86 |
| 262-758 | 0.31 | 0.18 | 0.29 |
| 759-1047 | 0.39 | 0.13 | 0.13 |
| 1048-1170 | 0.25 | 0.51 | 0.76 |
| 1171-1235 | 0.41 | 0.58 | 0.51 |
| 1236-1549 | 0.29 | 0.02 | 0.38 |
| 1550-2355 | 0.41 | 0.02 | 0.04 |
| 2356-2539 | 0.24 | 0.002 | 0.13 |

Table A.5: Estimated $d$ values for the regimes together with the p-value reported by the Beran (1992) test for $d$ as estimated for both the series and the regimes of the General Motors stock.

| Period | d | d=0.36 | d=d(t) |
|---|---|---|---|
| 1-2539 | 0.36 | 0.09 | - |
| 1-1515 | 0.27 | 0.20 | 0.76 |
| 1516-1875 | 0.34 | 0.60 | 0.65 |
| 1876-1956 | 0.09 | 0.06 | 0.61 |
| 1957-2006 | 0.29 | 0.29 | 0.43 |
| 2007-2151 | 0.25 | 0.39 | 0.60 |
| 2152-2269 | 0.34 | 0.46 | 0.47 |
| 2270-2423 | 0.19 | 0.86 | 0.93 |
| 2424-2539 | 0.13 | 0.05 | 0.51 |

Table A.6: Estimated $d$ values for the regimes together with the p-value reported by the Beran (1992) test for $d$ as estimated for both the series and the regimes of the Hewlett-Packard stock.

| Period | d | d=0.44 | d=d(t) |
|---|---|---|---|
| 1-2539 | 0.44 | 0.002 | - |
| 1-344 | 0.34 | 0.05 | 0.27 |
| 345-983 | 0.34 | 0.15 | 0.48 |
| 984-1535 | 0.39 | 0.11 | 0.29 |
| 1536-1723 | 0.40 | 0.75 | 0.80 |
| 1724-2006 | 0.29 | 0.20 | 0.76 |
| 2007-2349 | 0.38 | 0.04 | 0.19 |
| 2350-2539 | 0.38 | 0.78 | 0.86 |

Table A.7: Estimated $d$ values for the regimes together with the p-value reported by the Beran (1992) test for $d$ as estimated for both the series and the regimes of the IBM stock.

| Period | d | d=0.44 | d=d(t) |
|---|---|---|---|
| 1-2539 | 0.44 | 8.9e-06 | - |
| 1-103 | 0.37 | 0.01 | 0.01 |
| 104-526 | 0.29 | 0.19 | 0.31 |
| 527-1168 | 0.38 | 0.09 | 0.20 |
| 1169-1461 | 0.41 | 0.58 | 0.62 |
| 1462-1723 | 0.39 | 0.03 | 0.10 |
| 1724-1863 | 0.35 | 0.54 | 0.65 |
| 1864-2146 | 0.38 | 0.03 | 0.07 |
| 2147-2261 | 0.44 | 0.80 | 0.79 |
| 2262-2355 | 0.05 | 0.20 | 0.83 |
| 2356-2478 | 0.27 | 0.57 | 0.88 |
| 2479-2539 | 0.10 | 0.21 | 0.47 |

Table A.8: Estimated $d$ values for the regimes together with the p-value reported by the Beran (1992) test for $d$ as estimated for both the series and the regimes of the Intel stock.

| Period | d | d=0.46 | d=d(t) |
|---|---|---|---|
| 1-2539 | 0.46 | $3.59 \times 10^{-7}$ | - |
| 1-740 | 0.33 | 0.003 | 0.06 |
| 741-1168 | 0.46 | 0.67 | 0.74 |
| 1169-1535 | 0.40 | 0.26 | 0.40 |
| 1536-2255 | 0.47 | 0.004 | 0.002 |
| 2256-2342 | 0.19 | 0.09 | 0.19 |
| 2343-2539 | 0.42 | 0.49 | 0.59 |

Table A.9: Estimated $d$ values for the regimes together with the p-value reported by the Beran (1992) test for $d$ as estimated for both the series and the regimes of the Johnson and Johnson stock.

| Period | d | d=0.40 | d=d(t) |
|---|---|---|---|
| 1-2539 | 0.40 | 0.16 | - |
| 1-165 | 0.25 | 0.32 | 0.69 |
| 166-638 | 0.30 | 0.40 | 0.55 |
| 639-990 | 0.32 | 0.48 | 0.64 |
| 991-1169 | 0.27 | 0.02 | 0.33 |
| 1170-1226 | 0.38 | 0.78 | 0.78 |
| 1227-1309 | 0.08 | 0.15 | 0.66 |
| 1310-1537 | 0.39 | 0.30 | 0.31 |
| 1538-1854 | 0.38 | 0.95 | 0.95 |
| 1855-2153 | 0.30 | 0.17 | 0.26 |
| 2154-2267 | 0.37 | 0.30 | 0.29 |
| 2268-2395 | 0.33 | 0.42 | 0.56 |
| 2396-2539 | 0.27 | 0.32 | 0.77 |

Table A.10: Estimated $d$ values for the regimes together with the p-value reported by the Beran (1992) test for $d$ as estimated for both the series and the regimes of the Coca-Cola stock.

| Period | d Est | d=0.42 | d=d(t) |
|---|---|---|---|
| 1-2539 | 0.42 | $2.75 \times 10^{-6}$ | - |
| 1-206 | 0.25 | 0.002 | 0.17 |
| 207-616 | 0.30 | 0.18 | 0.54 |
| 617-1026 | 0.32 | 0.009 | 0.07 |
| 1027-1171 | 0.29 | 0.86 | 0.96 |
| 1172-1231 | 0.43 | 0.59 | 0.59 |
| 1232-1565 | 0.27 | 0.25 | 0.64 |
| 1566-1879 | 0.31 | 0.45 | 0.72 |
| 1880-2002 | 0.15 | 0.33 | 0.78 |
| 2003-2156 | 0.21 | 0.58 | 0.86 |
| 2157-2262 | 0.34 | 0.13 | 0.29 |
| 2263-2390 | 0.19 | 0.01 | 0.52 |
| 2391-2539 | 0.21 | 0.39 | 0.84 |

Table A.11: Estimated $d$ values for the regimes together with the p-value reported by the Beran (1992) test for $d$ as estimated for both the series and the regimes of the Merck stock.

| Period | d | d=0.39 | d=d(t) |
|---|---|---|---|
| 1-2539 | 0.39 | 0.34 | - |
| 1-252 | 0.27 | 0.03 | 0.22 |
| 253-1168 | 0.35 | 0.36 | 0.48 |
| 1169-1519 | 0.35 | 0.04 | 0.11 |
| 1520-1854 | 0.29 | 0.36 | 0.62 |
| 1855-2154 | 0.28 | 0.74 | 0.90 |
| 2155-2244 | 0.34 | 0.45 | 0.57 |
| 2245-2390 | 0.21 | 0.61 | 0.79 |
| 2391-2539 | 0.30 | 0.70 | 0.78 |

Table A.12: Estimated $d$ values for the regimes together with the p-value reported by the Beran (1992) test for $d$ as estimated for both the series and the regimes of the Microsoft stock.

| Period | d | d=0.46 | d=d(t) |
|---|---|---|---|
| 1-2539 | 0.46 | 0.008 | - |
| 1-333 | 0.25 | $1.84 \times 10^{-5}$ | 0.11 |
| 334-674 | 0.33 | 0.07 | 0.50 |
| 675-759 | 0.39 | 0.62 | 0.67 |
| 760-922 | 0.45 | 0.59 | 0.59 |
| 923-1105 | 0.44 | 0.34 | 0.35 |
| 1106-1535 | 0.38 | 0.29 | 0.51 |
| 1536-1620 | 0.30 | 0.62 | 0.74 |
| 1621-1720 | 0.40 | 0.66 | 0.74 |
| 1721-1854 | 0.45 | 0.41 | 0.41 |
| 1855-2360 | 0.44 | 0.86 | 0.89 |
| 2361-2539 | 0.41 | 0.67 | 0.77 |

Table A.13: Estimated $d$ values for the regimes together with the p-value reported by the Beran (1992) test for $d$ as estimated for both the series and the regimes of the Pfizer stock.

| Period | d | d=0.42 | d=d(t) |
|---|---|---|---|
| 1-2539 | 0.42 | $4.43 \times 10^{-6}$ | - |
| 1-144 | 0.30 | 0.59 | 0.70 |
| 145-399 | 0.23 | 0.49 | 0.95 |
| 400-485 | 0.35 | 0.27 | 0.37 |
| 486-898 | 0.34 | 0.45 | 0.62 |
| 899-1537 | 0.39 | 0.78 | 0.85 |
| 1538-1797 | 0.27 | 0.85 | 0.94 |
| 1798-1862 | 0.31 | 0.51 | 0.61 |
| 1863-2128 | 0.30 | 0.66 | 0.89 |
| 2129-2262 | 0.35 | 0.46 | 0.56 |
| 2263-2437 | 0.23 | 0.36 | 0.82 |
| 2438-2539 | 0.14 | 0.05 | 0.40 |

Table A.14: Estimated $d$ values for the regimes together with the p-value reported by the Beran (1992) test for $d$ as estimated for both the series and the regimes of the Walmart stock.

| Period | d Est | $d = 0.42$ | $d = d(t)$ |
|---|---|---|---|
| 1-2538 | 0.42 | $2.95 \times 10^{-12}$ | - |
| 1-289 | 0.26 | 0.008 | 0.03 |
| 290-1055 | 0.37 | 0.78 | 0.91 |
| 1056-1170 | 0.13 | 0.22 | 0.79 |
| 1171-1226 | 0.19 | 0.27 | 0.39 |
| 1227-1549 | 0.35 | 0.25 | 0.52 |
| 1550-1862 | 0.30 | 0.02 | 0.21 |
| 1863-1996 | 0.30 | 0.05 | 0.31 |
| 1997-2140 | 0.30 | 0.86 | 0.95 |
| 2141-2256 | 0.41 | 0.47 | 0.49 |
| 2257-2390 | 0.21 | 0.41 | 0.86 |
| 2391-2539 | 0.35 | 0.17 | 0.39 |

Table A.15: Estimated $d$ values for the regimes together with the p-value reported by the Beran (1992) test for $d$ as estimated for both the series and the regimes of the Exxon stock.

| Period | d | d=0.44 | d=d(t) |
|---|---|---|---|
| 1-2539 | 0.44 | 0.09 | - |
| 1-783 | 0.33 | 0.04 | 0.55 |
| 784-1552 | 0.37 | 0.13 | 0.37 |
| 1553-1614 | 0.37 | 0.91 | 0.92 |
| 1615-2165 | 0.37 | 0.02 | 0.05 |
| 2166-2260 | 0.35 | 0.20 | 0.38 |
| 2261-2349 | 0.35 | 0.48 | 0.51 |
| 2350-2539 | 0.17 | 0.003 | 0.31 |

## A.2 Graphical $d$ estimate with Regime Length Distributions

In this section all the results of the Graphical method outlined in Section (8.2) are presented. Most use an older style of presentation than in the main body of the thesis in which a number of individual dots are presented from the simulated data. The dots give a visual guide to the unconditional distribution while the lines are the conditional distribution of the $d$ estimate given regime length, as discussed in the main body of the thesis.

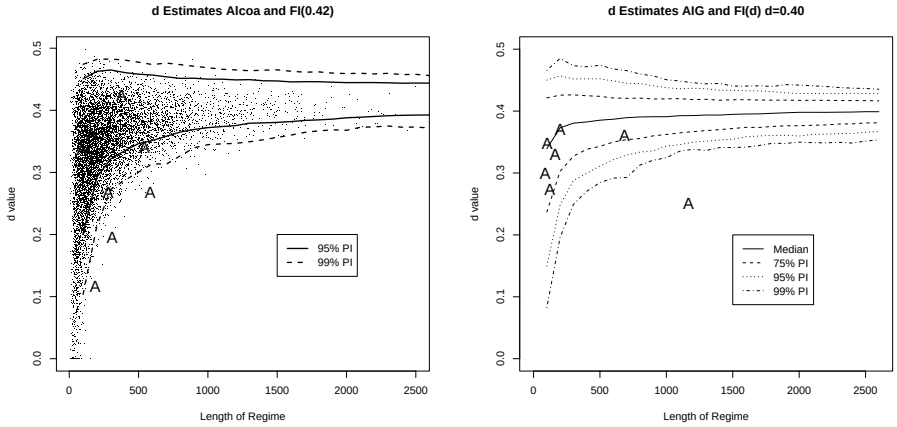


Figure A.1: The distribution of d values of the regimes for the Alcoa (left panel) and American International Group (right panel) stocks.

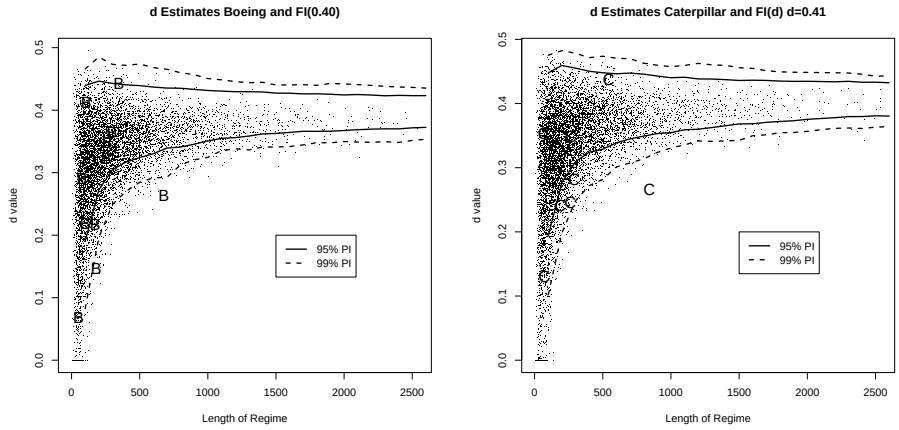


Figure A.2: The distribution of d values of the regimes for the Boeing (left panel) and Caterpillar (right panel) stocks.

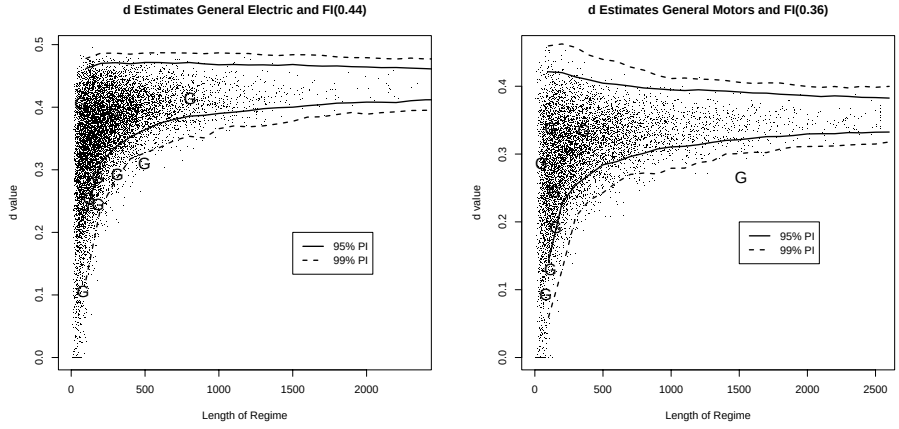


Figure A.3: The distribution of d values of the regimes for the General Electric (left panel) and General Motors (right panel) stocks.

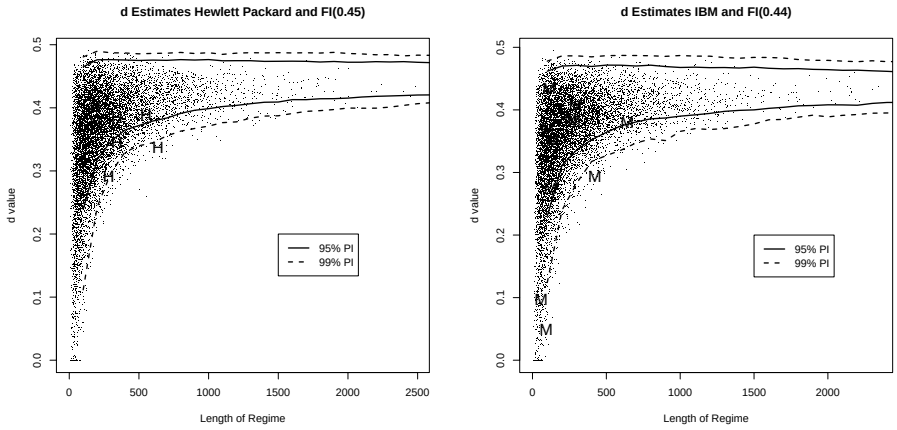


Figure A.4: The distribution of d values of the regimes for the Hewlett-Packard (left panel) IBM (right panel) stocks.

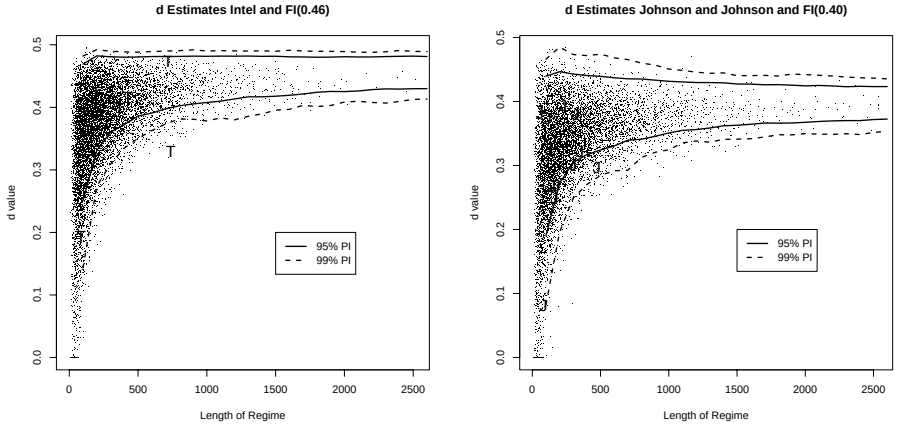


Figure A.5: The distribution of d values of the regimes for the Intel (left panel) Johnson and Johnson (right panel) stocks.

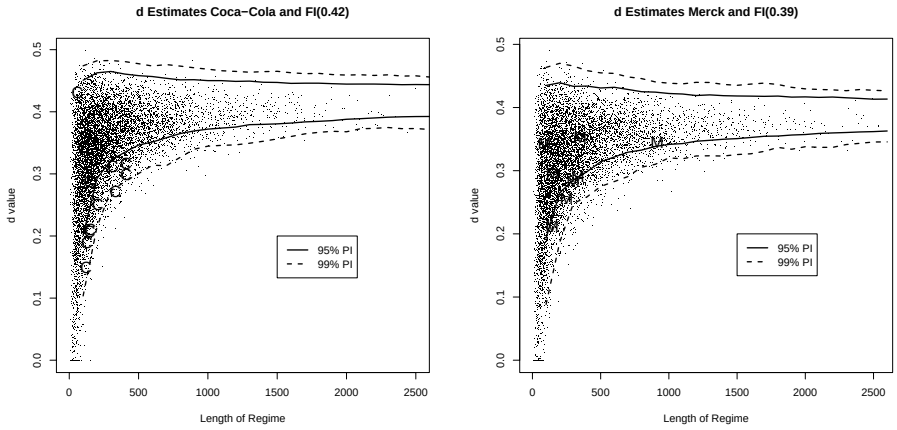


Figure A.6: The distribution of d values of the regimes for the Coca-Cola (left panel) Merck (right panel) stocks.

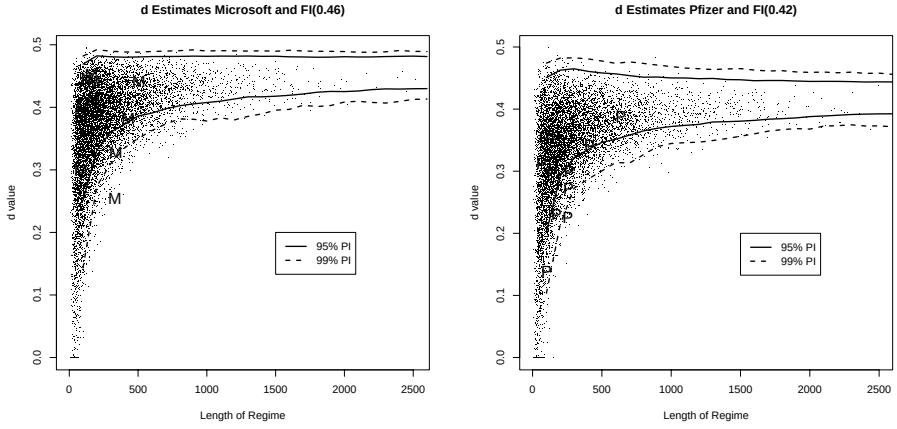


Figure A.7: The distribution of d values of the regimes for the Microsoft (left panel) Pfizer (right panel) stocks.

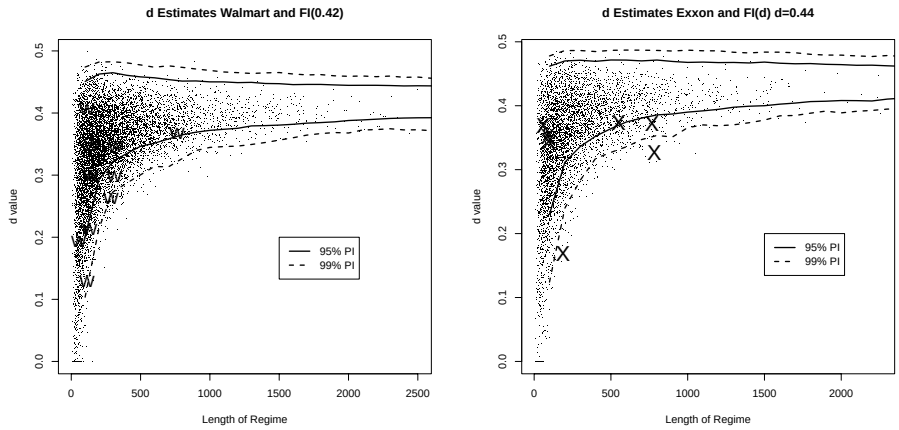


Figure A.8: The distribution of d values of the regimes for the Wal-Mart (left panel) Exxon (right panel) stocks.

## A.3 Bivariate Standard Deviation and Regime Length Distributions

This section presents all the results of the Graphical method outlined in Section (8.2), for the standard deviations. Most use an older style of presentation than that used in the main body of the thesis in which a number of individual dots are presented from the simulated data which give a visual guide to the unconditional distribution. The lines are the conditional distribution of the standard deviations, given regime length.

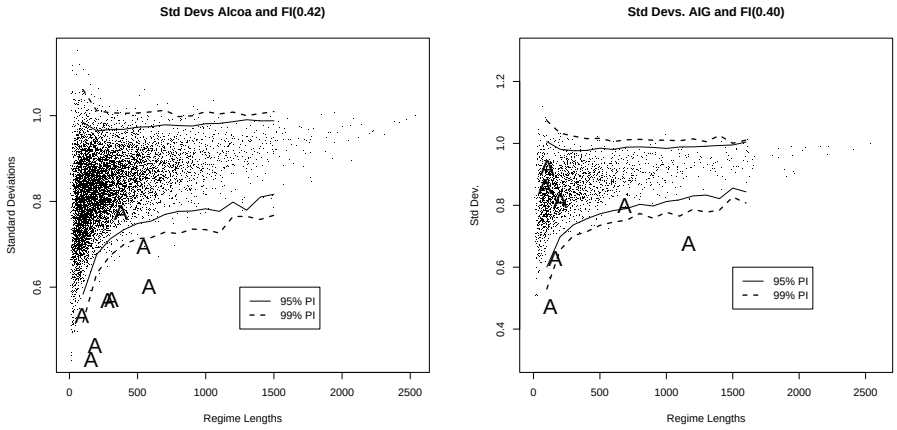


Figure A.9: The distribution of standard deviations of the regimes for the Alcoa (left panel) and American International Group (right panel) stocks.

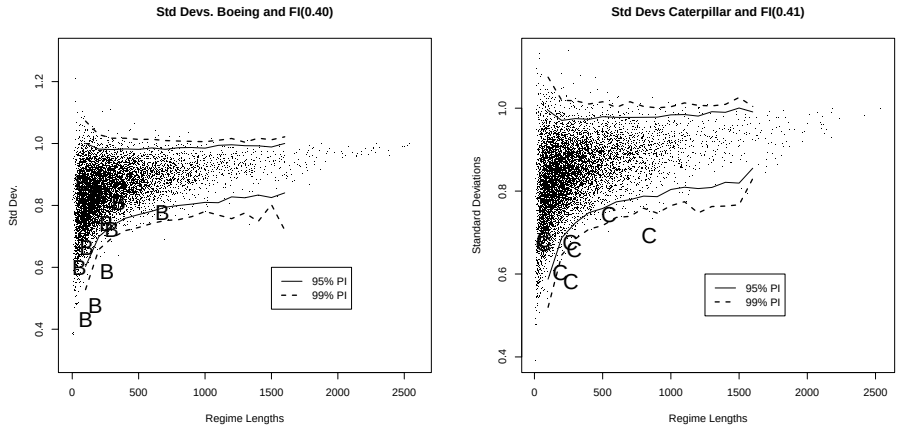


Figure A.10: The distribution of standard deviations of the regimes for the Boeing (left panel) and Caterpillar (right panel) stocks.

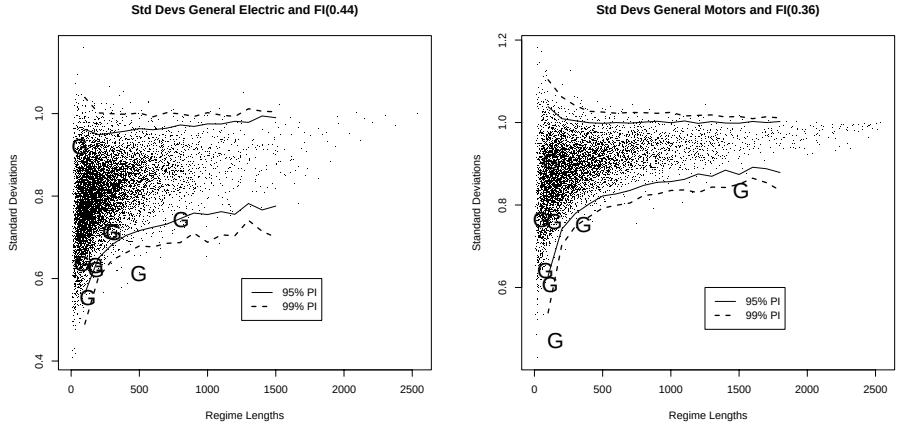


Figure A.11: The distribution of standard deviations of the regimes for the General Electric (left panel) and General Motors (right panel) stocks.

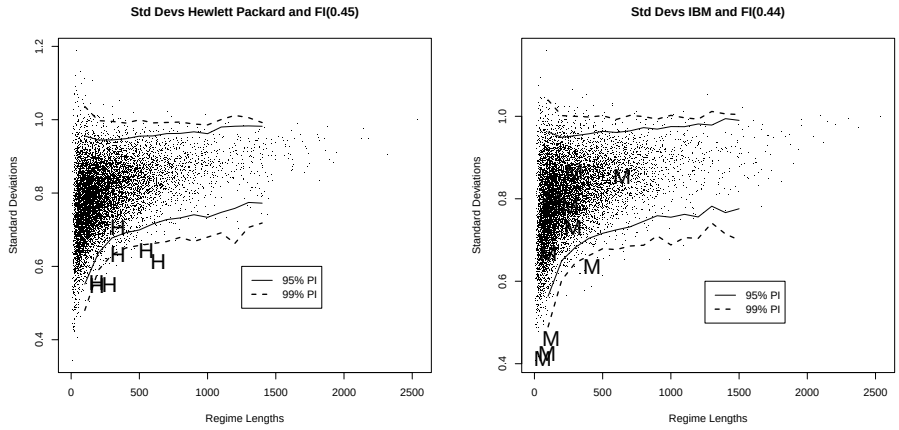


Figure A.12: The distribution of standard deviations of the regimes for the Hewlett-Packard (left panel) and IBM (right panel) stocks.

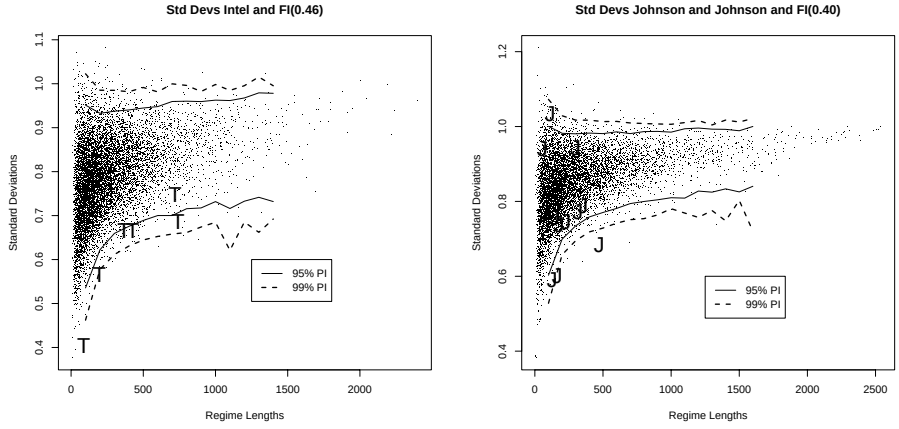


Figure A.13: The distribution of standard deviations of the regimes for the Intel (left panel) and Johnson and Johnson (right panel) stocks.

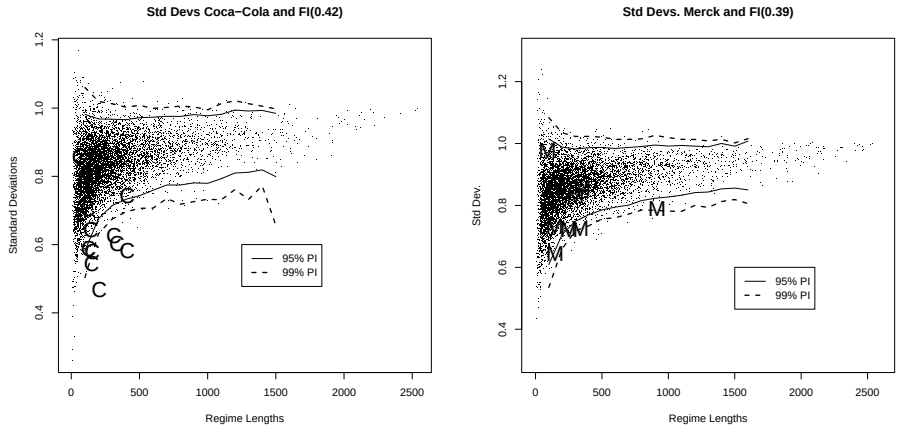


Figure A.14: The distribution of standard deviations of the regimes for the Coca-Cola (left panel) and Merck (right panel) stocks.

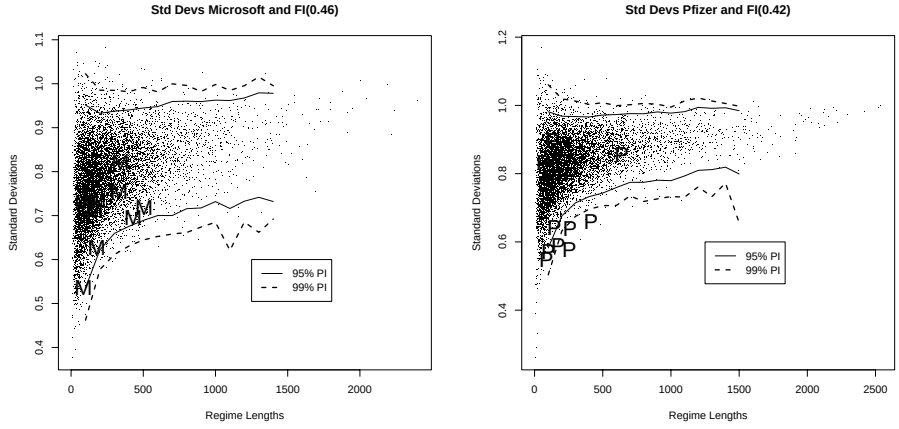


Figure A.15: The distribution of standard deviations of the regimes for the Microsoft (left panel) and Pfizer (right panel) stocks.

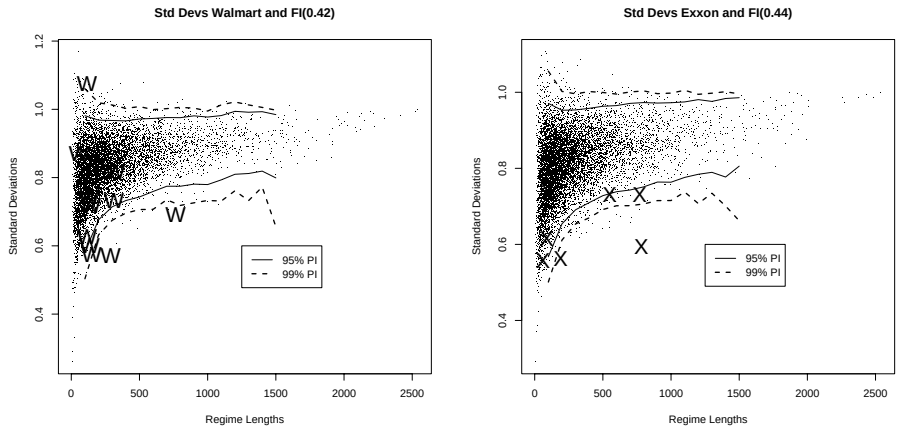


Figure A.16: The distribution of standard deviations of the regimes for the Wal-Mart (left panel) and Exxon (right panel) stocks.

## A.4 Break-CUSUM Results

In this section all of the results for the bivariate breaks-CUSUM distribution for the financial data are presented.

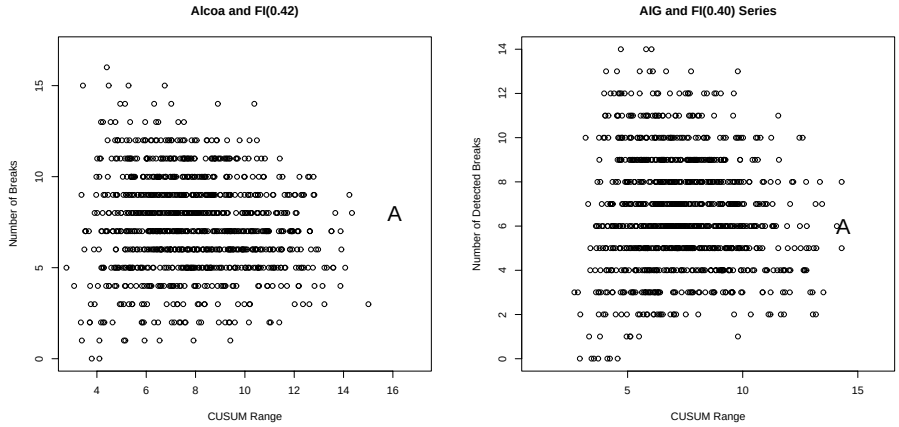


Figure A.17: The distribution of CUSUM ranges against numbers of reported breaks for the Alcoa (left panel) and American International Group (right panel) stocks.

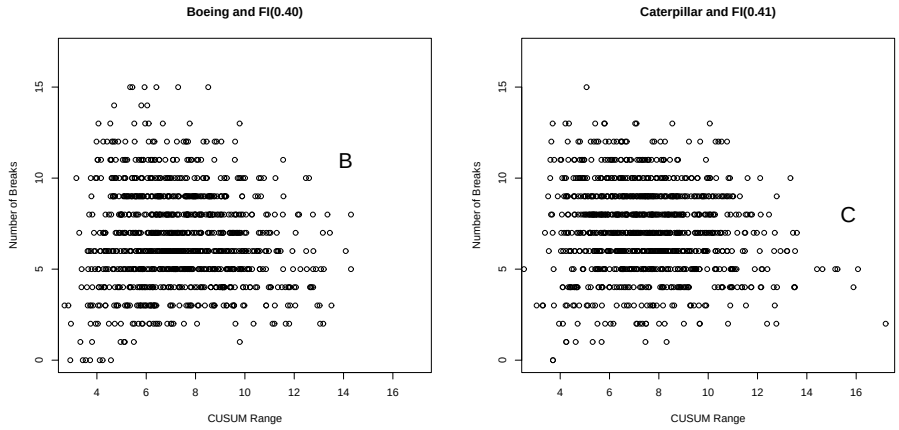


Figure A.18: The distribution of CUSUM ranges against numbers of reported breaks for the Boeing (left panel) and Caterpillar (right panel) stocks.

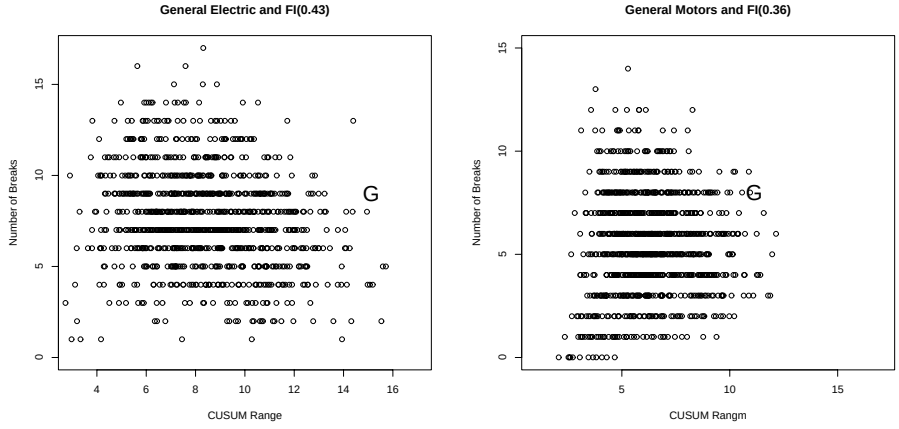


Figure A.19: The distribution of CUSUM ranges against numbers of reported breaks for the General Electric (left panel) and General Motors (right panel) stocks.

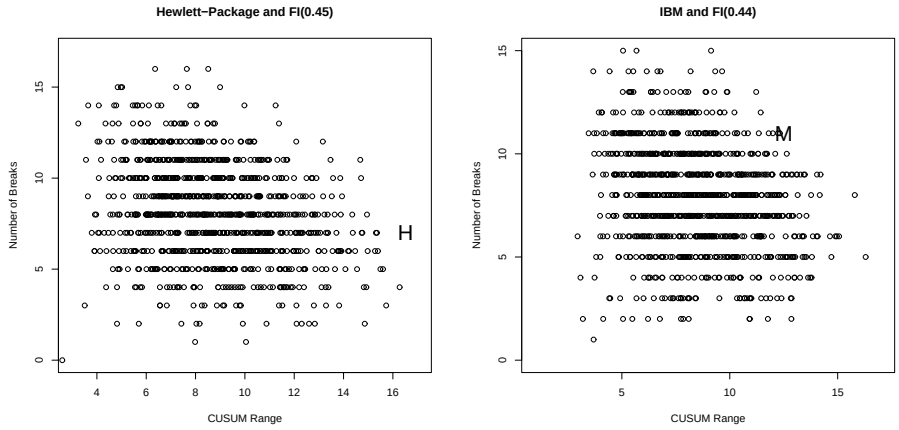


Figure A.20: The distribution of CUSUM ranges against numbers of reported breaks for the Hewlett-Packard (left panel) and IBM (right panel) stocks.

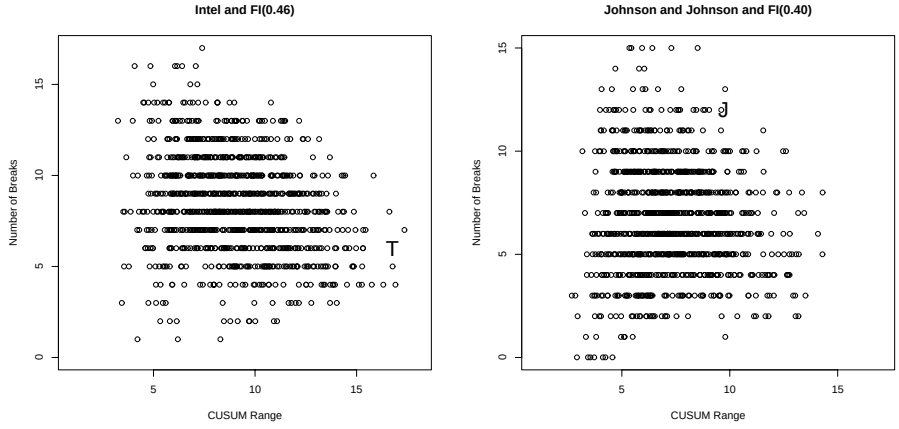


Figure A.21: The distribution of CUSUM ranges against numbers of reported breaks for the Intel (left panel) and Johnson and Johnson (right panel) stocks.

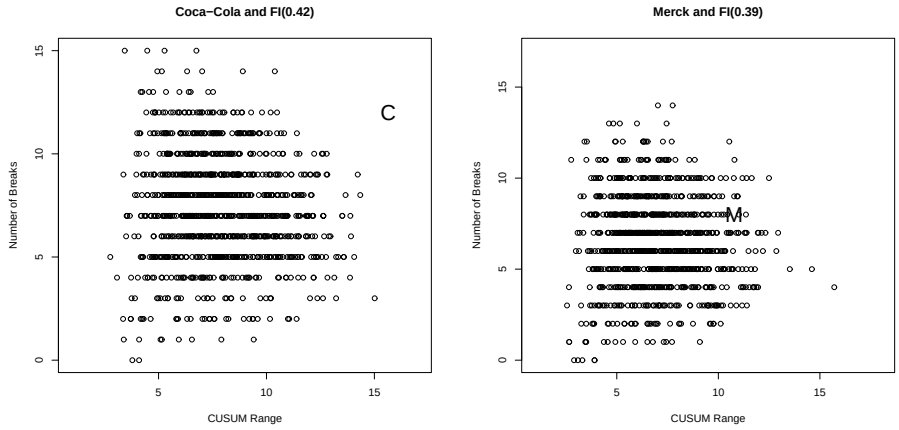


Figure A.22: The distribution of CUSUM ranges against numbers of reported breaks for the Coca-Cola (left panel) and Merck (right panel) stocks.

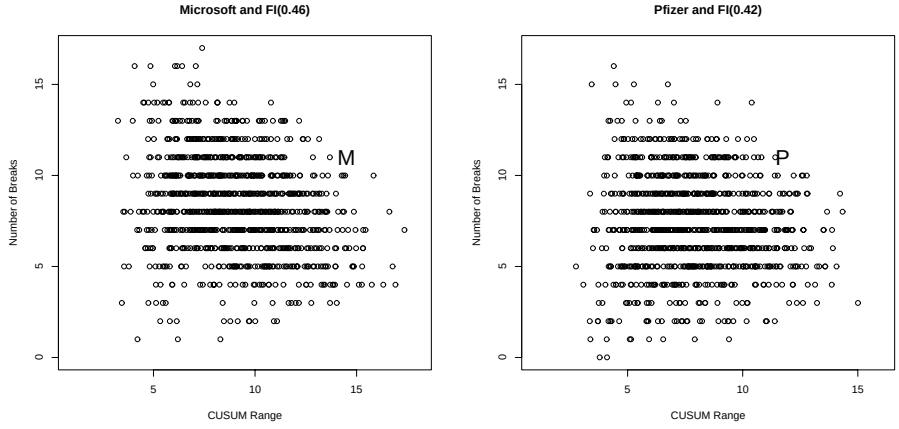


Figure A.23: The distribution of CUSUM ranges against numbers of reported breaks for the Microsoft (left panel) and Pfizer (right panel) stocks.

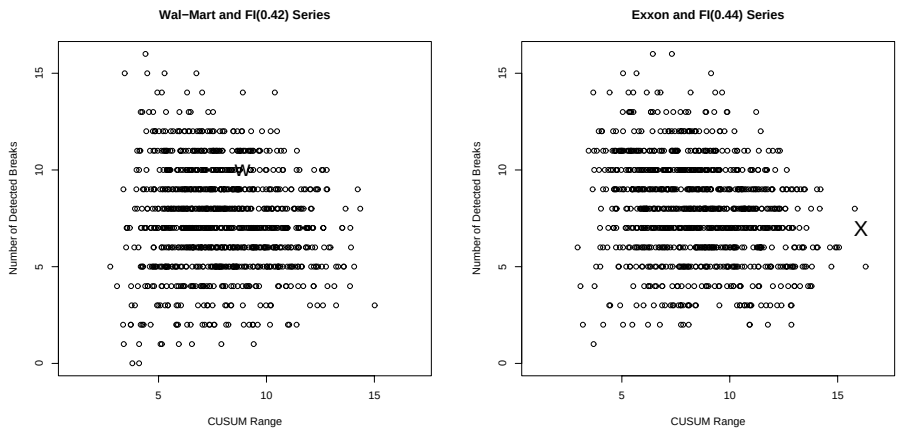


Figure A.24: The distribution of CUSUM ranges against numbers of reported breaks for the Wal-Mart (left panel) and Exxon (right panel) stocks.

# Appendix B

# Extra Geophysical Results

## B.1 Other Campito Mountain Results

Table B.1: Basic statistics for Campito Mountain Data for raw and first differenced series.

| Statistic | Value | First-Differenced |
|---|---|---|
| Mean | 42.29 | 0.004 |
| Standard Deviation | 11.77 | 9.21 |
| Skewness | 0.169 | -0.170 |
| Kurtosis | 0.458 | 1.677 |
| Jarque-Bera | | |
| p-value | $< 2.2 \times 10^{-16}$ | $< 2.2 \times 10^{-16}$ |

Table B.2: Basic statistics for Campito subseries reported by ART. The second line of the 2791 to 2574 BC period is the same data with two outliers adjusted to the mean value.

| Period | Mean | Std. Dev. | Skewness | Kurtosis | Jarque-Bera p-value |
|---|---|---|---|---|---|
| 3435 to 3284 BC | 58.15 | 11.75 | -0.069 | 0.448 | 0.435 |
| 3283 to 3100 BC | 46.60 | 7.32 | 0.010 | 0.250 | 0.731 |
| 3099 to 2792 BC | 36.49 | 7.66 | 0.139 | 0.176 | 0.473 |
| 2791 to 2574 BC | 54.89 | 9.86 | -0.777 | 4.342 | $2.2 \times 10^{-16}$ |
| | 55.08 | 9.16 | -0.056 | 0.942 | 0.0122 |
| 2573 to 1544 BC | 45.20 | 9.01 | -0.180 | 1.209 | $2.2 \times 10^{-16}$ |
| 1543 to 1216 BC | 52.00 | 9.96 | -0.376 | 0.896 | $6.058 \times 10^{-5}$ |
| 1215 to 359 BC | 35.39 | 7.93 | -0.124 | 1.104 | $7.913 \times 10^{-11}$ |
| 358 BC to 440 AD | 44.01 | 10.76 | -0.209 | 0.706 | $1.081 \times 10^{-5}$ |
| 441 to 862 AD | 36.59 | 10.61 | -0.287 | 0.131 | 0.045 |
| 863 to 1017 AD | 27.20 | 7.74 | 0.233 | -0.123 | 0.477 |
| 1018 to 1362 AD | 47.69 | 13.01 | 0.652 | 0.820 | $1.180 \times 10^{-7}$ |
| 1363 to 1862 AD | 36.31 | 9.67 | -0.069 | -0.137 | 0.6807 |
| 1863 to 1969 AD | 53.28 | 10.36 | -0.259 | 0.013 | 0.5335 |

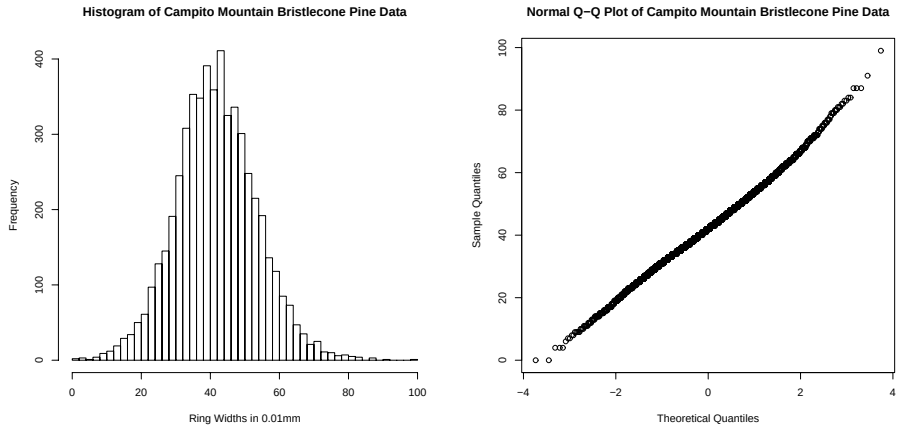


Figure B.1: Left panel Histogram the Campito Mountain ring thicknesses. Right panel Q-Q plot of the same.

## B.2 $H$ Estimates and Beran Tests

Table B.3: $H$ estimates and Beran (1992) p-values for the data sets considered for the study of long memory in temperature reconstruction series.

| Data Set | $H$ Estimate | Beran p-value |
|---|---|---|
| Baffin Island | 0.68 | 0.22 |
| Burgundy | 0.61 | 0.76 |
| Colorado | 0.95 | 0.44 |
| Moberg | 0.99 | 0 |
| Northern Hemisphere STD | 0.913 | 0.356 |
| Northern Hemisphere RCS | 0.947 | 0.495 |
| Shihua | 0.84 | 0.38 |
| Torneträsk | 0.84 | 0.17 |
| Urals | 0.64 | $4.46 \times 10^{-4}$ |
| West Greenland | 0.73 | $4.51 \times 10^{-4}$ |
| Western USA | 0.75 | 0.61 |

Table B.4: $H$ and $d$ estimates for six temperature reconstructions reported by 11 different estimators. Colo - Colorado Plateau, Nth Hem. - Northern Hemisphere, Torne - Torneträsk, W USA - Western USA reconstruction of Mann et al. (1998).

| Estimator | Colo | Nth Hem. | Shihua | Tasmania | Torne | W USA |
|---|---|---|---|---|---|---|
| Abs. val. | 0.508 | 0.779 | 0.895 | 0.604 | 0.773 | 0.672 |
| Agg. var. | 0.488 | 0.724 | 0.858 | 0.592 | 0.740 | 0.639 |
| Box Per | 0.911 | 0.934 | 0.884 | 0.768 | 0.889 | 0.701 |
| Dif. var | 0.787 | 1.122 | 1.084 | 0.844 | 1.173 | 0.954 |
| Higuchi | 0.966 | 0.982 | 0.967 | 0.967 | 0.810 | 0.631 |
| Peng | 0.875 | 0.973 | 0.929 | 0.822 | 0.926 | 0.758 |
| Periodogram | 0.990 | 0.977 | 1.039 | 0.733 | 1.015 | 0.810 |
| R/S | 0.895 | 0.639 | 0.690 | 0.809 | 0.732 | 0.657 |
| Whittle | 0.955 | 0.949 | 0.838 | 0.997 | 0.842 | 0.740 |
| Wavelet | 0.729 | 0.998 | 0.887 | 0.875 | 1.014 | 0.643 |
| H-R | 0.498 | 0.494 | 0.393 | 0.499 | 0.412 | 0.294 |

## B.3 Elk Lake

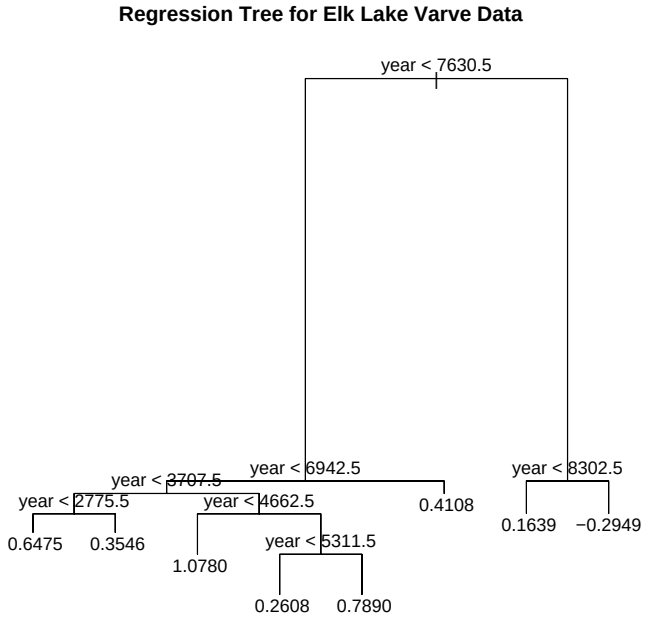


Figure B.2: Regression tree for the Elk Lake varve data.

Table B.5: H estimate for the regimes and the cumulative series.

| Date Range | Regime H | Cumulative H |
|---|---|---|
| 1-2775 | 0.742 | 0.742 |
| 2776-3707 | 0.786 | 0.767 |
| 3708-4662 | 0.815 | 0.794 |
| 4663-5311 | 0.818 | 0.809 |
| 5312-6942 | 0.886 | 0.831 |
| 6943-7630 | 0.889 | 0.841 |
| 7631-8302 | 0.902 | 0.848 |
| 8303-10224 | 0.811 | 0.860 |

Table B.6: Estimates of $H$ reported by the nine different estimators for Elk Lake and simulated Elk Lake data.

| Method | H Estimate | Std. Dev. | Simulated |
|---|---|---|---|
| Abs. Val. | 0.940 | 0.035 | 0.920 |
| Agg. Var. | 0.938 | 0.042 | 0.914 |
| Boxed Per. | 0.950 | 0.024 | 0.892 |
| Diff. Var. | 0.776 | 0.243 | 0.901 |
| Higuchi | 0.947 | 0.036 | 0.935 |
| Peng | 0.883 | 0.033 | 0.879 |
| Periodogram | 0.972 | 0.020 | 0.951 |
| R/S | 0.961 | 0.082 | 0.773 |
| Whittle | 0.860 | 0.143 | 0.853 |

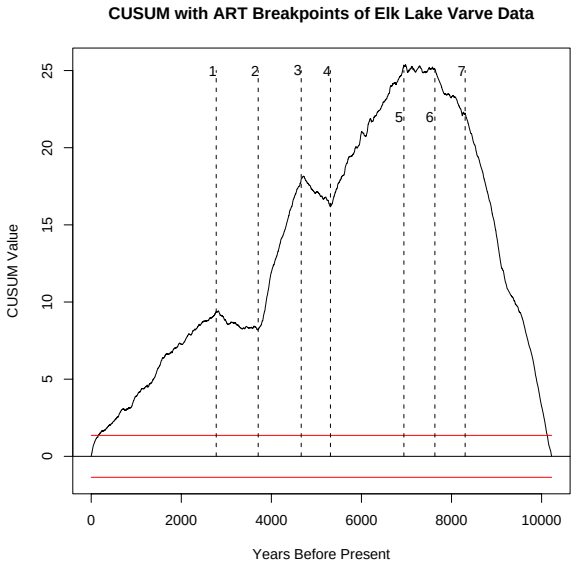


Figure B.3: CUSUM plot with ART break points for the Elk Lake varve data.

Table B.7: Breaks in the Elk Lake data which correspond to known climate or physical changes in the lake.

| Change Date (BP) | Change |
|---|---|
| 4662 | Drought (Forman et al., 2001) |
| 5311 | Chemical change in the lake water (Forman et al., 2001) |
| 8302 | Collapse of Laurentide ice sheet (Dean et al., 2002) |

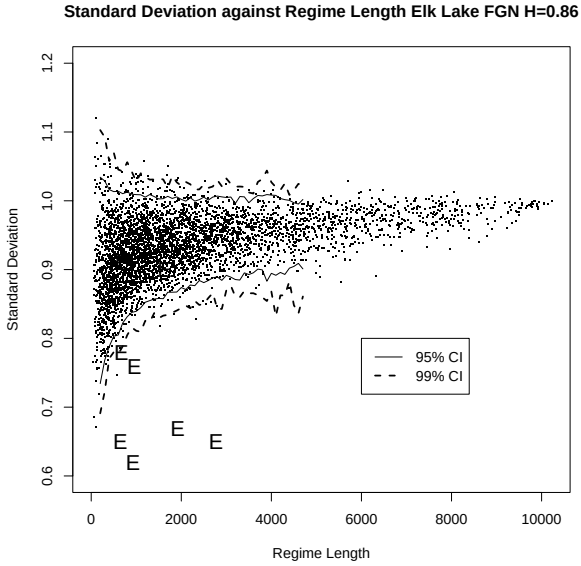


Figure B.4: Bivariate distribution of standard deviation with regime length for FGNs with $H = 0.86$. The "E" symbols are the Elk Lake varve data.

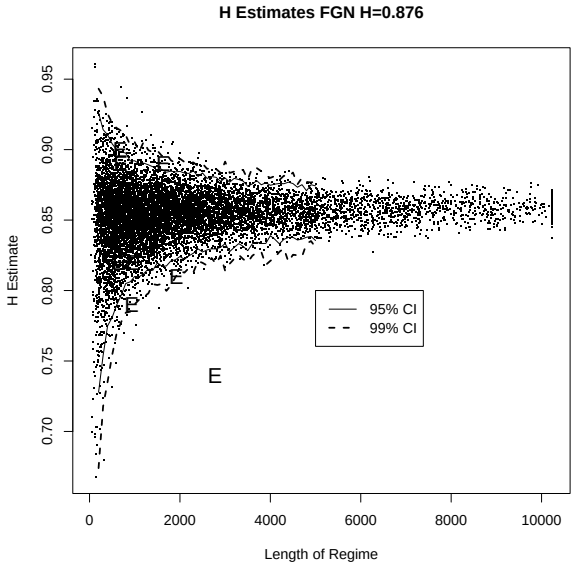


Figure B.5: Distribution of $H$ estimates for regime of various lengths. "E" is the Elk Lake data.

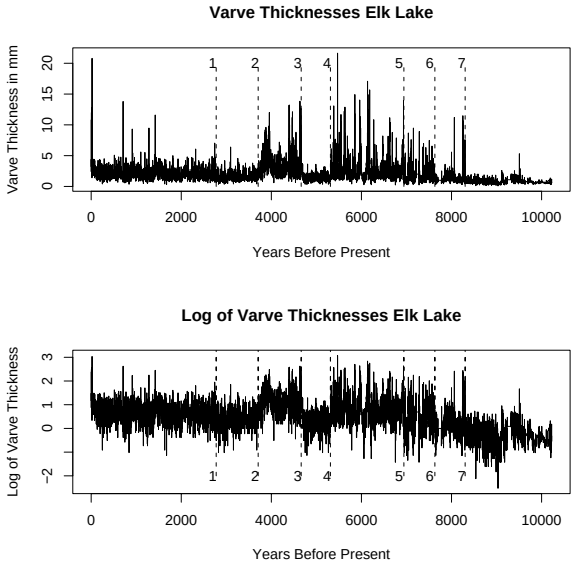


Figure B.6: Elk Lake data with log transform and ART breaks.

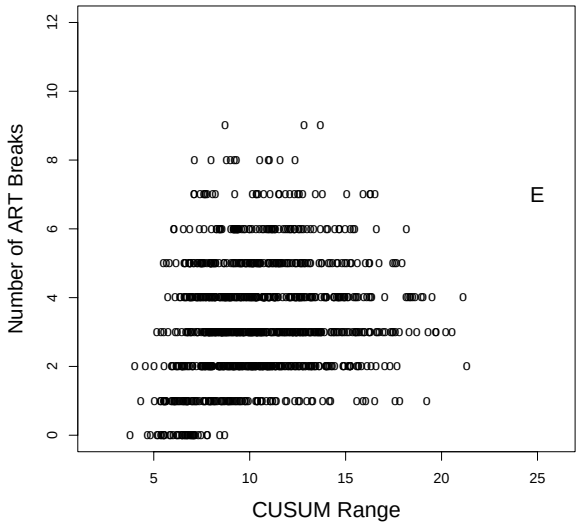


Figure B.7: CUSUM range against ART breaks for FGN with $H = 0.86$ with Elk Lake data plotted as an "E".

## B.4 Temperature Reconstruction Time Series

Table B.8: Basic statistics for the Torneträsk reconstruction and the regimes reported by ART.

| Period | Mean | Std Dev. | Skew | Kurt | Jarque-Bera |
|---|---|---|---|---|---|
| 1-1992 | -0.153 | 0.902 | 0.018 | 0.065 | 0.784 |
| 1-491 | -0.005 | 0.838 | -0.037 | -0.427 | 0.160 |
| 492-535 | -0.459 | 0.589 | -0.084 | -0.978 | 0.476 |
| 536-554 | -1.970 | 0.772 | -0.240 | -1.296 | 0.555 |
| 555-962 | -0.363 | 0.886 | -0.122 | -0.352 | 0.227 |
| 963-1107 | 0.686 | 0.819 | 0.231 | -0.396 | 0.350 |
| 1108-1192 | -0.119 | 0.740 | -0.884 | 0.818 | ***0.001 |
| 1193-1354 | -0.661 | 0.779 | -0.126 | -0.463 | 0.429 |
| 1355-1567 | 0.179 | 0.852 | -0.115 | -0.150 | 0.739 |
| 1568-1751 | -0.885 | 0.746 | -0.211 | 0.387 | 0.251 |
| 1752-1914 | -0.261 | 0.931 | -0.265 | -0.355 | 0.266 |
| 1915-1992 | 1.144 | 0.863 | 0.154 | -0.821 | 0.330 |

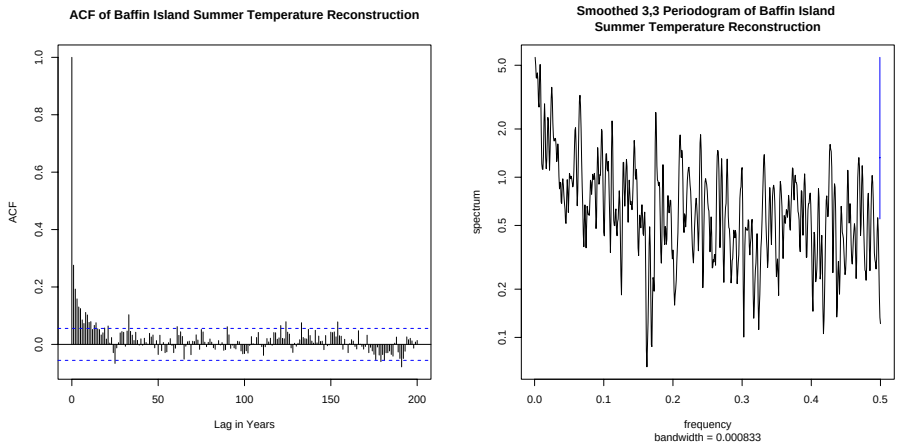


Figure B.8: ACF and estimated spectrum of temperature reconstruction for Baffin Island of Moore et al. (2001).

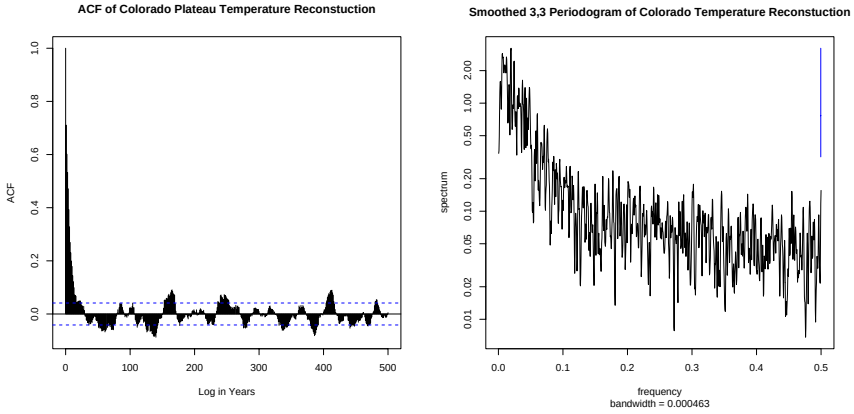


Figure B.9: ACF and estimated spectrum of temperature reconstruction for Colorado Plateau of Salzer and Kipfmueller (2005).

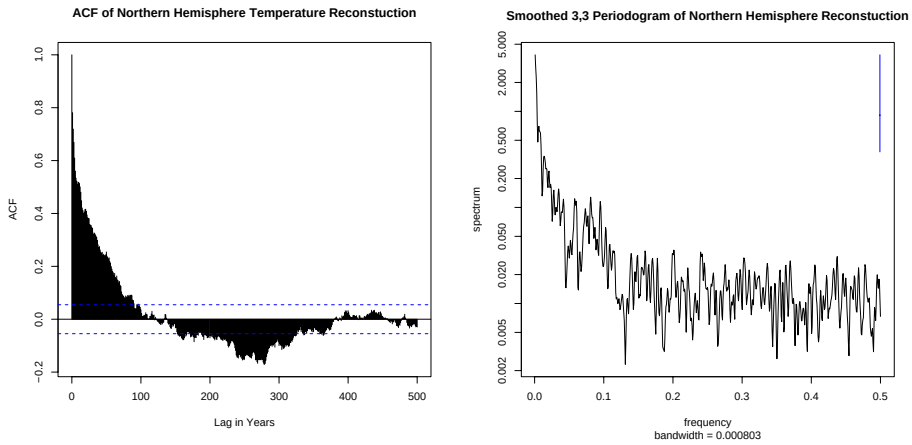


Figure B.10: ACF and spectral estimate of the temperature reconstruction for the Northern Hemisphere of D'Arrigo et al. (2006).

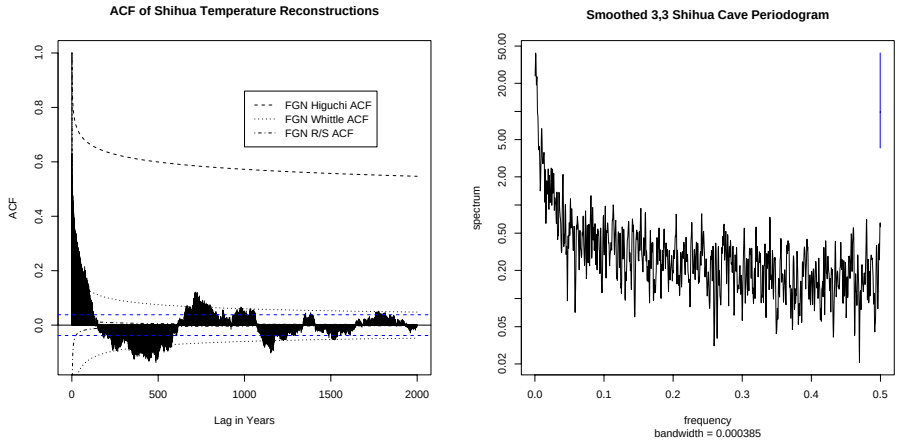


Figure B.11: ACF and spectral estimate of temperature reconstruction for Shihua Cave of Tan et al. (2001).

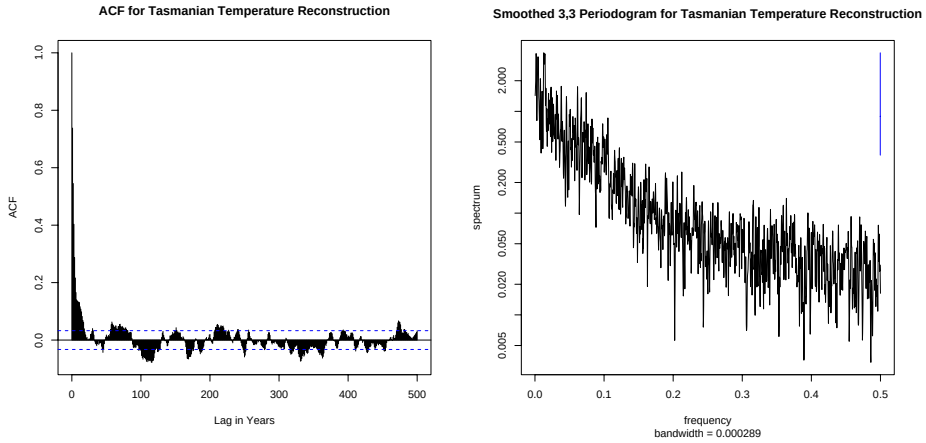


Figure B.12: ACF and spectral estimate of temperature reconstruction for Tasmania of Cook et al. (2000).

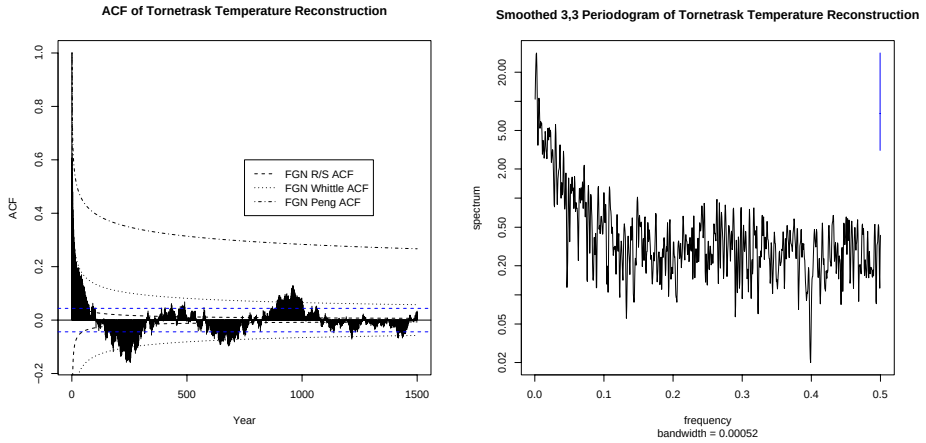


Figure B.13: ACF and estimated spectrum of temperature reconstruction for Torneträsk of Briffa et al. (1992).

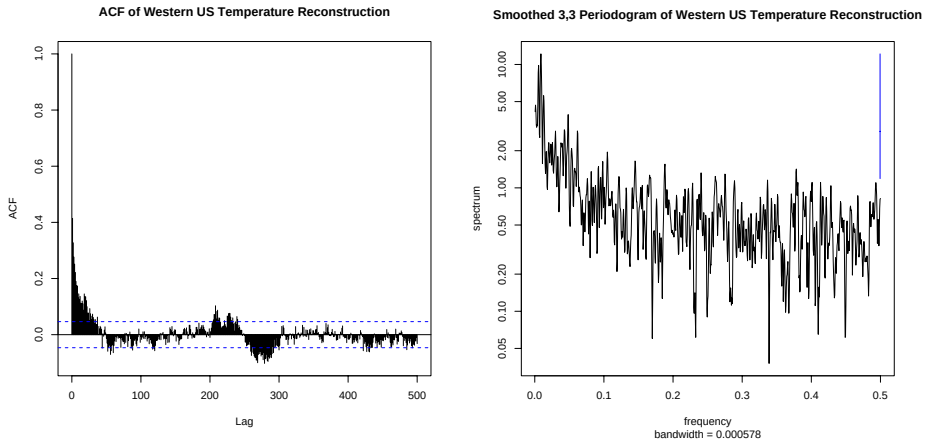


Figure B.14: ACF and estimated spectrum of temperature reconstruction for Western USA of Mann et al. (1998).

Table B.9: Beran (1992) p-values for constant and variable H values for the Baffin Island temperature reconstruction.

| Period | $H$ | $H = 0.683$ | H=H(t) |
|---|---|---|---|
| 752 to 1992 | 0.683 | 0.22 | - |
| 752 to 1372 | 0.688 | 0.87 | 0.88 |
| 1373 to 1480 | 0.517 | 0.02 | 0.19 |
| 1481 to 1543 | 0.612 | 0.37 | 0.34 |
| 1544 to 1975 | 0.647 | 0.09 | 0.09 |
| 1976 to 1992 | 0.740 | 0.72 | 0.72 |

Table B.10: $H$ Estimates and p-values for the Beran (1992) goodness-of-fit test for the Northern Hemisphere STD series of D'Arrigo et al. (2006) and its regimes.

| Period | $H$ Est | $H = 0.913$ | $H = H(t)$ |
|---|---|---|---|
| 713-1995 | 0.913 | 0.356 | - |
| 713-740 | 0.884 | 0.819 | 0.806 |
| 741-850 | 0.907 | 0.002 | 0.002 |
| 851-871 | 0.805 | 0.234 | 0.157 |
| 872-1107 | 0.824 | 0.863 | 0.903 |
| 1108-1207 | 0.751 | 0.371 | 0.674 |
| 1208-1811 | 0.876 | 0.542 | 0.617 |
| 1812-1850 | 0.876 | 0.430 | 0.389 |
| 1851-1922 | 0.628 | 0.205 | 0.565 |
| 1923-1995 | 0.884 | 0.817 | 0.840 |

Table B.11: $H$ Estimates and p-values for the Beran (1992) goodness-of-fit test for the RCS Northern Hemisphere series of D'Arrigo et al. (2006) and its regimes.

| Period | $H$ Est | $H$=0.947 | $H = H(t)$ |
|---|---|---|---|
| 713-1995 | 0.947 | 0.495 | - |
| 713-724 | 0.758 | 0.771 | 0.762 |
| 725-926 | 0.930 | 0.783 | 0.762 |
| 927-1119 | 0.816 | 0.879 | 0.905 |
| 1120-1856 | 0.889 | 0.897 | 0.912 |
| 1857-1919 | 0.660 | 0.385 | 0.751 |
| 1920-1995 | 0.839 | 0.857 | 0.886 |

Table B.12: H Estimates and p-values returned by the Beran (1992) goodness-of-fit test for the Shihua Cave temperature reconstruction.

| Estimator | H or d Estimate | Beran p-value |
|---|---|---|
| Absolute value | 0.895 | 0.287 |
| Aggregated variance | 0.858 | 0.432 |
| Boxed Periodogram | 0.884 | 0.350 |
| Differenced Variance | 1.084 | - |
| Higuchi | 0.967 | 0.0054 |
| Peng | 0.929 | 0.0856 |
| Periodogram | 1.039 | - |
| Rescaled Range | 0.690 | 0 |
| Whittle | 0.838 | 0.377 |
| Wavelet | 0.887 | 0.333 |
| Haslett-Raftery | (d=0.393) 0.893 | 0.297 |

Table B.13: $H$ Estimates and p-values for the Beran (1992) goodness-of-fit test for the Shihua Cave series and regimes.

| Period | $H$ Est. | $H$=0.84 | $H = H(t)$ |
|---|---|---|---|
| 1-2549 | 0.84 | 0.38 | - |
| 1-158 | 0.63 | 0.39 | 0.85 |
| 159-598 | 0.68 | 0.24 | 0.65 |
| 599-1123 | 0.81 | 0.56 | 0.49 |
| 1124-1190 | 0.80 | 0.57 | 0.49 |
| 1191-1447 | 0.82 | 0.80 | 0.80 |
| 1448-1511 | 0.69 | 0.81 | 0.90 |
| 1512-1608 | 0.67 | 0.46 | 0.65 |
| 1609-1863 | 0.66 | 0.64 | 0.76 |
| 1864-2114 | 0.78 | 0.88 | 0.88 |
| 2115-2245 | 0.87 | 0.78 | 0.80 |
| 2246-2451 | 0.84 | 0.91 | 0.91 |
| 2452-2549 | 0.94 | 0.10 | 0.26 |

Table B.14: $H$ estimates reported by 11 estimators and their Beran $p$-values for the Torneträsk reconstruction.

| Estimator | H Estimate | Beran p-value |
|---|---|---|
| Absolute value | 0.773 | 0.0007 |
| Aggregated variance | 0.740 | $1.146 \times 10^{-10}$ |
| Boxed Periodogram | 0.889 | 0.089 |
| Differenced Variance | 1.173 | - |
| Higuchi | 0.810 | 0.075 |
| Peng | 0.926 | 0.015 |
| Periodogram | 1.016 | - |
| Rescaled Range | 0.732 | $3.542 \times 10^{-14}$ |
| Whittle | 0.843 | 0.166 |
| Wavelet | 1.014 | - |
| Haslett-Raftery | ($d$=0.412) 0.912 | 0.034 |

Table B.15: Estimated $H$ values and Beran (1992) test p-values for the Torneträsk temperature reconstruction.

| Period | $H$ est. | $H$=0.84 | $H = H(t)$ |
|---|---|---|---|
| 1-1992 | 0.843 | 0.166 | - |
| 1-491 | 0.80 | 0.22 | 0.19 |
| 492-535 | 0.56 | 0.02 | 0.41 |
| 536-554 | 0.70 | 0.85 | 0.85 |
| 555-962 | 0.82 | 0.65 | 0.61 |
| 963-1107 | 0.77 | 0.73 | 0.54 |
| 1108-1192 | 0.61 | 0.81 | 0.87 |
| 1193-1354 | 0.75 | 0.05 | 0.16 |
| 1355-1567 | 0.81 | $2.70 \times 10^{-4}$ | $1.81 \times 10^{-4}$ |
| 1568-1751 | 0.70 | 0.39 | 0.62 |
| 1752-1914 | 0.83 | 0.44 | 0.43 |
| 1915-1992 | 0.82 | 0.76 | 0.77 |

Table B.16: $H$ Estimates and p-values for the Beran (1992) goodness-of-fit test for the Western USA series and regimes.

| Period | $H$ Est. | $H$=0.74 | $H = H(t)$ |
|---|---|---|---|
| 200-1980 | 0.740 | 0.613 | - |
| 200-1319 | 0.709 | 0.726 | 0.846 |
| 1320-1334 | 0.485 | 0.406 | 0.558 |
| 1335-1382 | 0.708 | 0.252 | 0.309 |
| 1383-1437 | 0.745 | 0.623 | 0.632 |
| 1438-1926 | 0.744 | 0.594 | 0.589 |
| 1927-1980 | 0.694 | 0.701 | 0.756 |

## B.5 Higuchi Estimator Results

The Higuchi (1988) estimator is somewhat different to other estimators of $H$ in that it estimates the fractal dimension, $D$, of a time series. As implemented in the contributed `R` package `fSeries` it reports an estimate of $H$ on the assumption that the series being analysed is an FGN. If the series is indeed an FGN then there is a simple arthimetic relationship (Taqqu et al., 1995, $D$=2-$H$) between $H$ and $D$. Unlike other estimators, which may return a value of $H$ which means little or nothing if the series is not an FGN, the Higuchi estimator still returns a useful estimate of the time series' fractal dimension, though in the `R` package `fSeries` one needs to convert $H$ back to $D$.

In time series analysis the analyst looks for quantities which are stationary across the whole series. As the claim that $H$ is stationary is refuted when using the Whittle estimator we analysed these series with the Higuchi estimator. We used rolling 500 year estimates of $H$ (as a proxy for $D$) for these reconstructions and compared them with rolling 500 data point estimates of $H$ in computer generated FGNs. The results for the computer generated FGNs are presented in Figure (B.15).

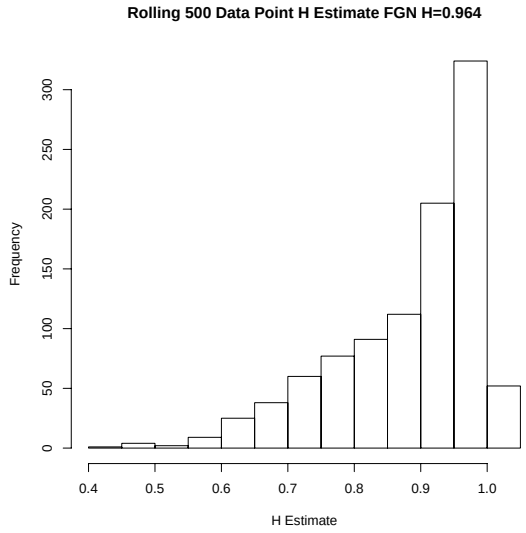


Figure B.15: Histogram of rolling 500 data point estimates of $H$ by the Higuchi (1988) estimator simulated FGN with H=0.964.

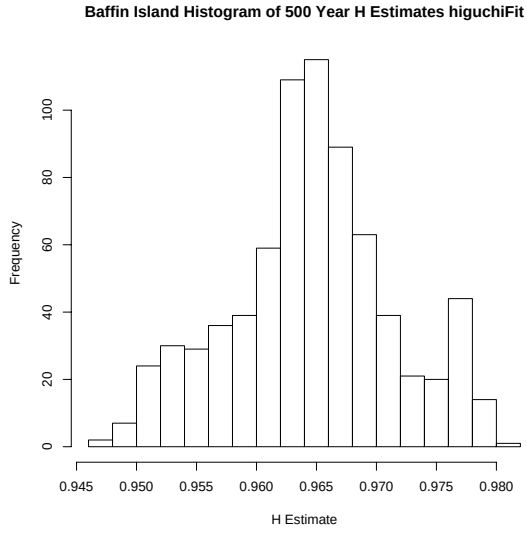


Figure B.16: Histogram of 500 year estimates of $H$ by the Higuchi (1988) estimator for the temperature reconstruction for Baffin Island of Moore et al. (2001).

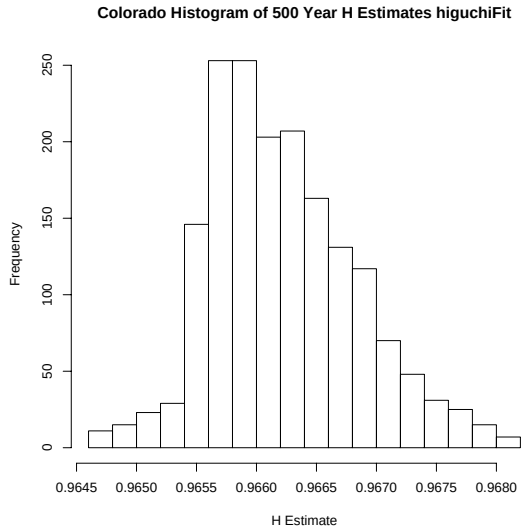


Figure B.17: Histogram of 500 year estimates of $H$ by the Higuchi (1988) estimator for the temperature reconstruction for the Colorado Plateau of Salzer and Kipfmueller (2005).

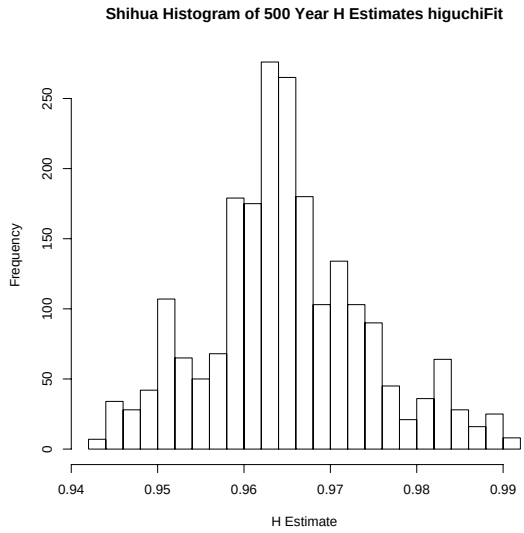


Figure B.18: Histogram of 500 year estimates of $H$ by the Higuchi (1988) estimator for the temperature reconstruction for the Shihua Cave of Tan et al. (2001).

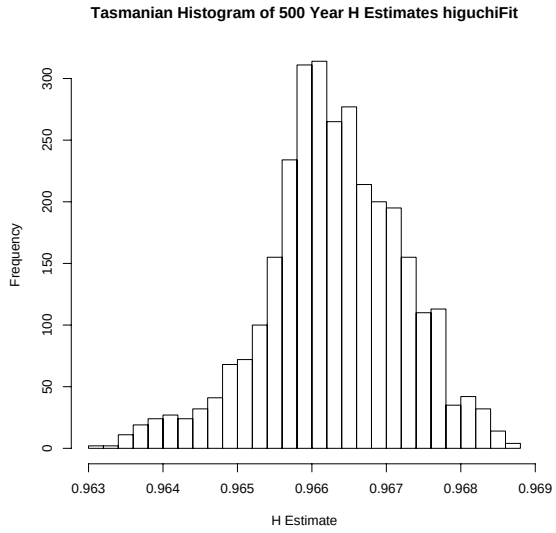


Figure B.19: Histogram of 500 year estimates of $H$ by the Higuchi (1988) estimator for the temperature reconstruction for Tasmania of Cook et al. (2000).

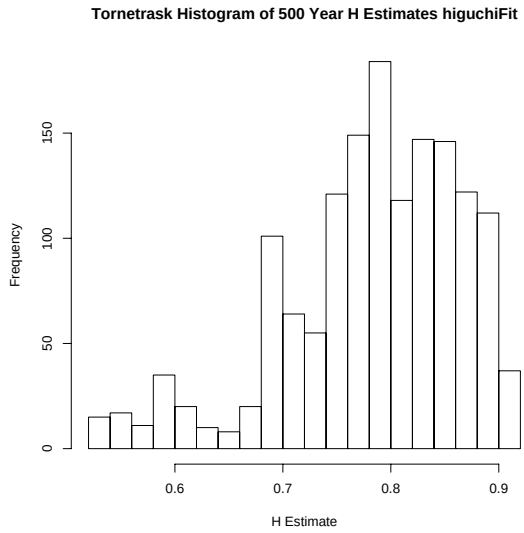


Figure B.20: Histogram of 500 year estimates of $H$ by the Higuchi (1988) estimator for the temperature reconstruction for Torneträsk of Briffa et al. (1992).

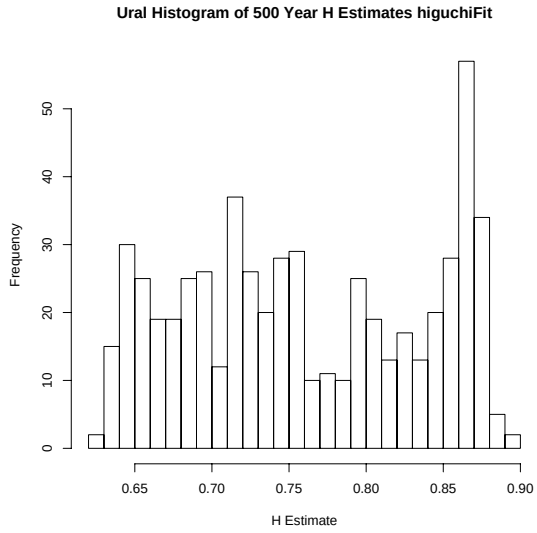


Figure B.21: Histogram of 500 year estimates of $H$ by the Higuchi (1988) estimator for the temperature reconstruction for the Urals of Briffa et al. (1995).

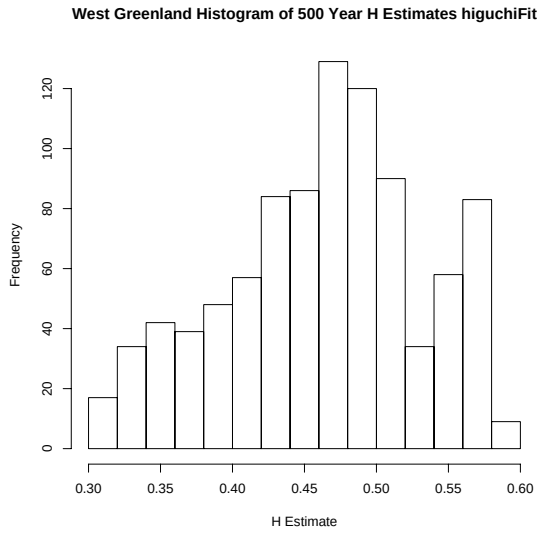


Figure B.22: Histogram of 500 year estimates of $H$ by the Higuchi (1988) estimator for the temperature reconstruction for the West Greenland of Fisher et al. (1996).

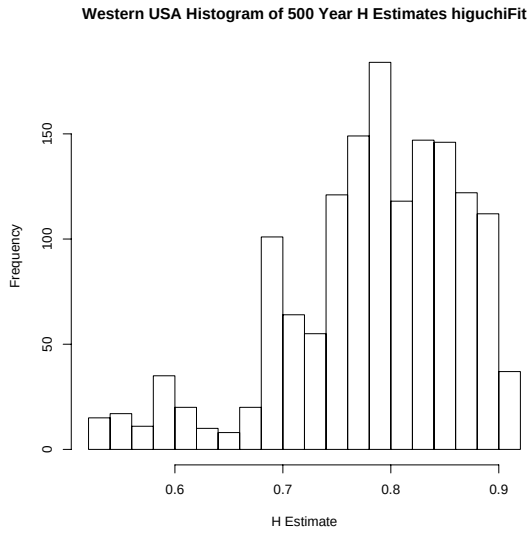


Figure B.23: Histogram of 500 year estimates of $H$ by the Higuchi (1988) estimator for the temperature reconstruction for the Western USA of Mann et al. (1998).

## B.6 Rolling $H$ Values

In a number of cases the breakpoints reported by ART also correspond to marked shifts in the $H$ estimate as reported by the Whittle estimator.

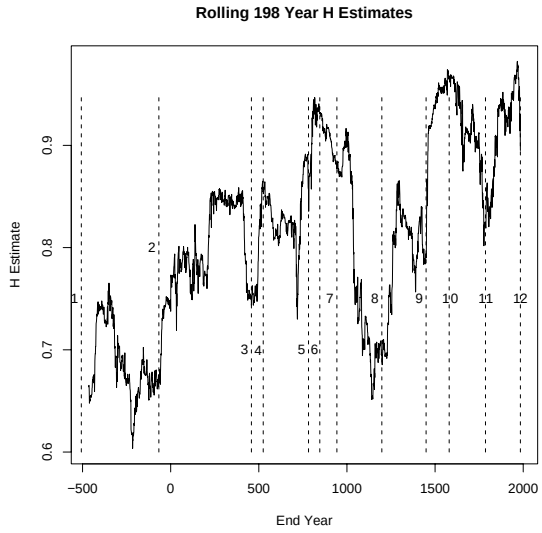


Figure B.24: Rolling 198 year $H$ estimates for the Shihua Cave warm season temperature reconstructions with the breaks reported by ART marked by dashed vertical lines.

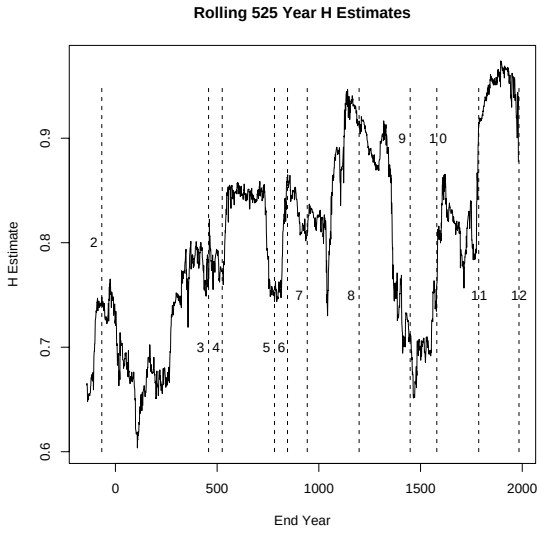


Figure B.25: Rolling 525 year $H$ estimates for the Shihua Cave warm season temperature reconstructions with the breaks reported by ART marked by dashed vertical lines.

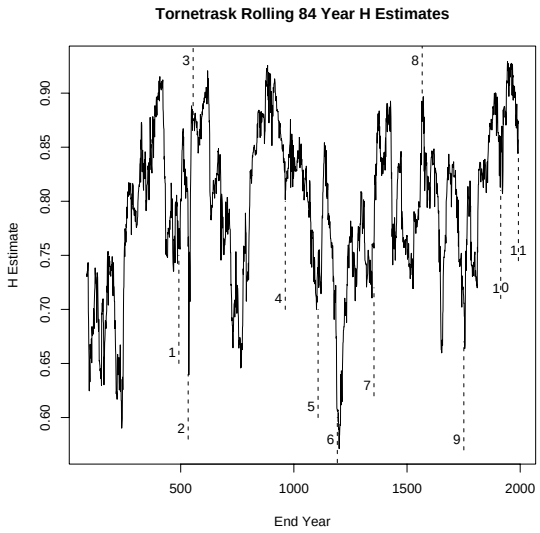


Figure B.26: Rolling 84 year $H$ estimates for the Torneträsk temperature reconstructions with the breaks reported by ART marked by dashed vertical lines.

## B.7 $H$ Distributions

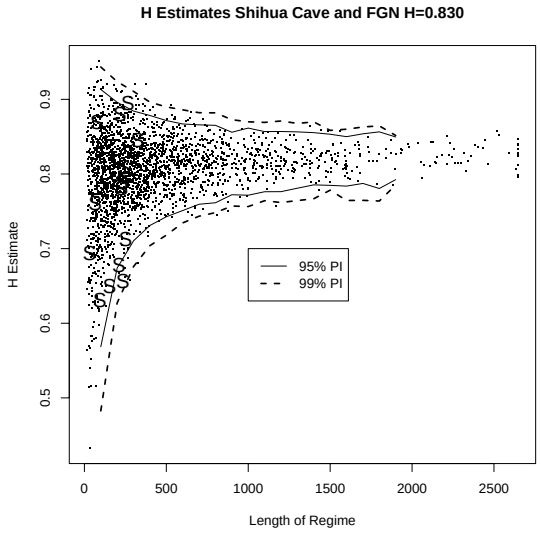


Figure B.27: Bivariate distribution of $H$ with regime length for Shihua Cave speleothems.

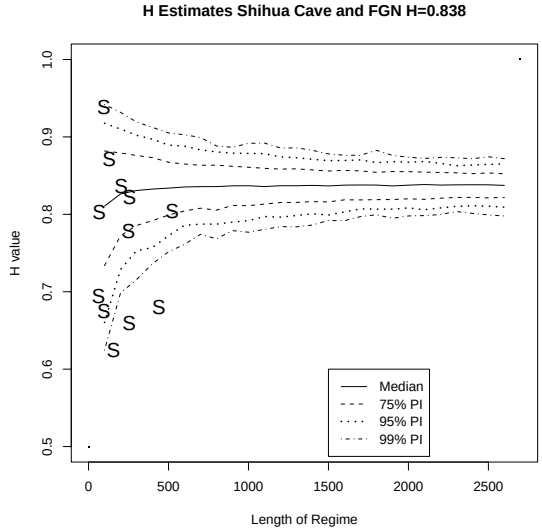


Figure B.28: Bivariate distribution of $H$ with regime length for Shihua Cave temperature reconstructions.

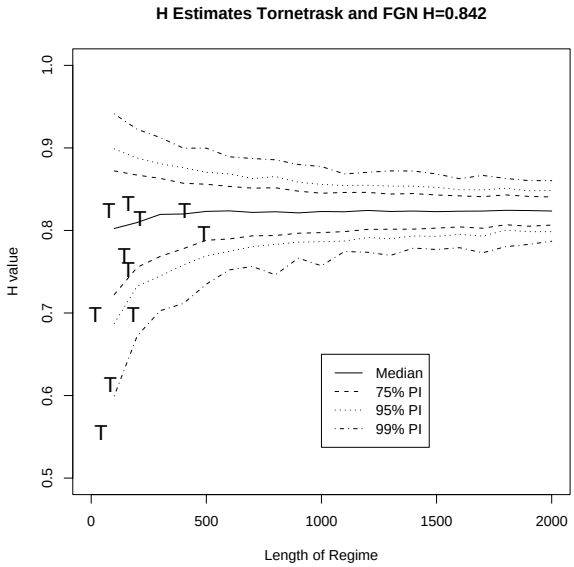


Figure B.29: Bivariate distribution of $H$ with regime length for Torneträsk temperature reconstructions.

# Bibliography


Abry, P. and Veitch, D. (1998). Wavelet Analysis of Long-Range-Dependent Traffic. *IEEE Transactions on Information Theory*, 44(1):2–15.

Abry, P., Veitch, D., and Flandrin, P. (1998). Long-Range Dependence: Revisiting Aggregation with Wavelets. *Journal of Time Series Analysis*, 19(3):253–266.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csaki, F., editors, *2nd International Symposium on Information Theory*, pages 267–281. Akademia Kadio, Budapest.

Alonso, A. M., Pena, D., and Romo, J. (2002). Forecasting time series with sieve bootstrap. *Journal of Statistical Planning and Inference*, 100:1–11.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, 61:43–76.

Andrews, D. W. K. (1993). Tests for Parameter Instability and Structural Change with Unknown Change Point. *Econometrica*, 61(4):821–856.

Andrews, D. W. K. and Ploberger, W. (1994). Optimal Tests when a Nuisance Parameter is Present Only Under the Alternative. *Econometrica*, 62(6):1383–1414.

Bai, J. (1998). A Note on Spurious Break. *Econometric Theory*, 14:663–669.

Bai, J. and Perron, P. (1998). Estimating and Testing Linear Models with Multiple Structural Changes. *Econometrica*, 66(1):47–78.

Bai, J. and Perron, P. (2003). Computation and Analysis of Multiple Structural Change Models. *Journal of Applied Econometrics*, 18:1–22.

Baillie, R. T. (1996). Long memory processes and fractional integration in economics. *Journal of Economics*, 73:5–59.

Baillie, R. T. and Chung, S.-K. (2002). Modeling and forecasting from trend-stationary long memory models with applications to climatology. *International Journal of Forecasting*, 18:215–226.

Banerjee, A. and Urga, G. (2005). Modelling structural breaks, long memory and stock market volatility: an overview. *Journal of Econometrics*, 129:1–34.

Barnard, G. (1959). Control Charts and Stochastic Processes. *Journal of the Royal Statistical Society B*, 21(2):100–115.

Bell, R. and Roth, R. (1969). Curve fitting by segmented straight lines. *Journal of American Statistical Association*, 64:1079–1084.

Beran, J. (1989). A Test of Location for Data with Slowly Decaying Serial Correlations. *Biometrika*, 76(2):261–269.

Beran, J. (1992). A Goodness-of-Fit Test for Time Series with Long Range Dependence. *Journal of the Royal Statistical Society B*, 54(3):749–760.

Beran, J. (1994). *Statistics for Long Memory Processes*. Chapman & Hall/CRC Press.

Beran, J. and Terrin, N. (1996). Testing for a change in the long-memory parameter. *Biometrika*, 83(3):627–638.

Beran, J. and Terrin, N. (1999). Testing for a change in the long-memory parameter. *Biometrika*, 86(1):233.

Beran, J., Whitcher, B., and Maechler, M. (2006). *longmemo: Statistics for Long-Memory Processes (Jan Beran) – Data and Functions*. R package version 0.9-3.

Berkes, I., Horvath, L., and Kokoszka, P. (2006). On Discriminating Between Long-Range Dependence and Changes in Mean. *The Annals of Statistics*, 34(3):1140–1165.

Bhattacharya, R., Gupta, V. K., and Waymire, E. (1983). The Hurst effect under trends. *Journal of Applied Probability*, 20:649–662.

Bloomfield, P. (1992). Trends in Global Temperature. *Climatic Change*, 21:1–16.

Bloomfield, P. (2000). *Fourier Analysis of Time Series Second Edition.* Wiley-Interscience.

Bloomfield, P. and Nychka, D. (1992). Climate Spectra and Detecting Climate Change. *Climatic Change*, 21:275–287.

Blum, J. R. and Rosenblatt, J. I. (1972). *Probability and Statistics.* W.B. Saunders.

Bollerslev, T. and Mikkelson, H. (1996). Modeling and pricing long memory in stock marker volatility. *Journal of Econometrics*, 73:151–184.

Bollerslev, T. and Wright, J. H. (2000). Semiparametric estimation of long memory volatility dependencies: the role of high frequency data. *Journal of Econometrics*, 98:81–106.

Bond, G., Kromer, B., Beer, J., Muscheler, R., Evans, M. N., Showers, W., Hoffman, S., Lotti-Bond, R., Hajdas, I., and Bonani, G. (2001). Persistent Solar Influence on North Atlantic Climate During the Holocene. *Science*, 294:2130–2136.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1993). *Classification and Regression Trees.* Chapman & Hall/CRC.

Briffa, K. R., Jones, P. D., Schweingruber, F. H., Shiyatov, S. G., and Cook, E. R. (1995). Unusual twentieth-century summer warmth is a 1,000-year temperature record from Siberia. *Nature*, 376:156–159.

Briffa, K. R., Jones, P. D., Wigley, T. M. L., Picher, J. R., and Ballie, M. G. L. (1992). Fennoscandian summers from 500AD: Temperature changes on short and long timescales. *Climate Dynamics*, 7:111–119.

Brockwell, P. J. and Davis, R. A. (2006). *Time Series: Theory and Methods*. Springer, 2nd edition.

Brown, R. L., Durbin, J., and Evans, J. (1975). Techniques for Testing the Constancy of Regression Relationships over Time. *Journal of the Royal Statistical Society Series B*, 37(2):149–192.

Cajueiro, D. O. and Tabak, B. M. (2005). Testing for time-varying long-range dependence in volatility for emerging markets. *Physica A*, 346:577–588.

Cappelli, C., Penny, R. N., Rea, W. S., and Reale, M. (2008). Detecting multiple mean breaks at unknown points with Atheoretical Regression Trees. *Mathematics and Computers in Simulation*, 78(2-3):351–356.

Cappelli, C., Rea, W. S., and Reale, M. (2007). The Application of Regression Trees to the Detecting of Multiple Structural Breaks in the Mean of a Time Series. Research Report UCMSD 2007/4, Department of Mathematics and Statistics, University of Canterbury.

Cappelli, C. and Reale, M. (2005). Detecting Changes in Mean Levels with Atheoretical Regression Trees. Research Report UCMSD 2005/2, Department of Mathematics and Statistics, University of Canterbury.

Chan, N. H. (2002). *Time Series Applications to Finance*. Wiley Interscience.

Chatfield, C. (2004). *The Analysis of Time Series 6th Edition*. Chapman & Hall/CRC Press.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART Model Search. *Journal of the American Statistical Association*, 93(443):935–948.

Chow, G. C. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica*, 28(3):591–605.

Chuine, I., Yiou, P., Viovy, N., Seguin, B., Daux, V., and Ladurie, E. L. R. (2004). Grape ripening as a past climate indicator. *Nature*, 432(7015):289–290.

Cioczek-Georges, R. and Mandelbrot, B. B. (1995). A Class of Micropulses and Antipresistent Fractional Brownian Motion. *Stochastic Processes and their Applications*, 60:1–18.

Cook, E. R., Buckley, B. M., D'Arrigo, R. D., and Peterson, M. J. (2000). Warm-season temperatures since 1600BC reconstructed from Tasmanian tree rings and their relationship to large-scale sea surface temperature anomalies. *Climate Dynamics*, 16:79–91.

Cooper, S. J. (1998). Multiple Regimes in US Output Fluctuations. *Journal of Business and Economic Statistics*, 16(1):92–100.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. Chapman & Hall/CRC.

Csörgo, M. and Horvath, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley Series in Probability and Statistics. Wiley, New York.

da Rosa, J. C., Veiga, A., and Medeiros, M. C. (2008). Tree-Structured Smooth Transition Regression Models Based on CART Algorithm. *Journal of Computational Statistics and Data Analysis*, 52:2469–2488.

Dalhaus, R. (1989). Efficient parameter estimation for self-similar processes. *The Annals of Statistics*, 17(4):1749–1766.

D'Arrigo, R., Wilson, R., and Jacoby, G. (2006). On the long-term context for late twentieth century warming. *Journal of Geophysical Research*, 111. doi:10.1029/2005JD006352.

Dean, W. (1994). Elk Lake High Resolution Data USGS Open File Report 94-578.

Dean, W. E. (2006). Email personnal communication – 6 July 2006.

Dean, W. E., Forester, R. M., and Bradbury, J. P. (2002). Early Holocene change in atmospheric circulation in the Northern Great Plains: an upstream view of the 8.2ka cold event. *Quaternary Science Reviews*, 21:1763–1775.

Denison, D. G. T., Malick, B. K., and Smith, A. F. M. (1998). A Bayesian CART Algorithm. *Biometrika*, 85(2):363–377.

Diebold, F. X. and Inoue, A. (2001). Long Memory and Regime Switching. *Journal of Econometrics*, 105:131–159.

Diggle, P. J. (1990). *Time Series: A Biostatistical Introduction*. Clarendon Press Oxford.

Doukhan, P., Oppenheim, G., and Taqqu, M. (2003). *Theory and Applications of Long-Range Dependence*. Birkhaüser.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

Embrechts, P. and Maejima, M. (2002). *Selfsimilar Processes*. Princeton University Press.

Esposito, F., Melerba, D., and Semeraro, G. (1997). A Comparitive Analysis of Methods for Pruning Decision Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):476–491.

Ferreira, P. E. (1975). A Bayesian Analysis of a Switching Regression Model: Known Number of Regimes. *Journal of the American Statistical Association*, 70(350):370–374.

Fisher, D. A., Koerner, R. M., Kuiviner, K., Clausen, H. B., Johnsen, S. J., Steffensen, J.-P., Gundestrup, N., and Hammer, C. U. (1996). Intercomparison of ice core and

precipitation records from sites in Canada and Greenland over the last few centuries using EOF techniques. In Jones, P. D., Bradley, R. S., and Jouzel, J., editors, *Climatic Variations and Forcing Mechanisms of the Last 2000 Years, NATO ASI Series*, volume 41. Springer-Verlag.

Fisher, W. D. (1958). On Grouping for Maximum Homogeneity. *Journal of the American Statistical Association*, 53(284):789–798.

Forman, S. L., Oglesby, R., and Webb, R. S. (2001). Temporal and spatial patterns of Holocene dune activity on the Great Plains of North America: megadroughts and climate links. *Global and Planetary Change*, 29:1–29.

Fox, R. and Taqqu, M. S. (1986). Large-Sample Properties of Parameter Estimates for Strongly Dependent Stationary Gaussian Time Series. *The Annals of Statistics*, 14(2):517–532.

Fraley, C., Leisch, F., Maechler, M., Reisen, V., and Lemonte, A. (2006). *fracdiff: Fractionally differenced ARIMA aka ARFIMA(p,d,q) models.* R package version 1.3-0.

Gagliardini, P., Trojani, F., and Urga, G. (2005). Robust GMM tests for structural breaks. *Journal of Econometrics*, 129:139–182.

Gipp, M. R. (2001). Interpretation of climate dynamics from phase space portraits: Is the climate strange or just different? *Paleoceanography*, 16(4):335–351.

Giraitis, L., Robinson, P. M., and Surgailis, D. (1999). Variance-type estimation of long memory. *Stochastic Processes and their Applications*, 80:1–24.

Granger, C. W. J. (1966). The Typical Spectral Shape of an Economic Variable. *Econometrica*, 34(1):150–161.

Granger, C. W. J. (1980). Long Memory Relationships and the Aggregation of Dynamic Models. *Journal of Econometrics*, 14:227–238.

Granger, C. W. J. and Hyung, N. (2004). Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. *Journal of Empirical Finance*, 11:213–228.

Granger, C. W. J. and Joyeux, R. (1980). An Introduction to Long-range Time Series Models and Fractional Differencing. *Journal of Time Series Analysis*, 1:15–30.

Granger, C. W. J. and Morris, M. J. (1976). Time Series Modelling and Interpretation. *Journal of the Royal Statistical Society A*, 139(2):246–257.

Granger, C. W. J. and Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford University Press, Oxford.

Granger, C. W. J. and Teräsvirta, T. (1999). A simple nonlinear time series model with misleading linear properties. *Economics Letters*, 62:161–165.

Guthery, S. (1974). Partition regression. *Journal of American Statistical Association*, 69:945–947.

Hamilton, J. (1994). *Time Series Analysis*. Princeton University Press.

Hansen, B. (2001). The New Econometrics of Structural Change: Dating Breaks in the US Labor Productivity. *Journal of Economic Perspectives*, 15(4):117–128.

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50(4):1029–1054.

Hantemirov, R. M. and Shiyatov, S. G. (2002). A continuous multimillenial ring-width chronology in Yamal, northwestern Siberia. *The Holocene*, 12(6):717–726.

Hare, S. R. and Mantua, N. J. (2000). Empirical evidence for North Pacific regime shifts in 1977 and 1989. *Progress in Oceanography*, 47:103–145.

Hartigan, J. (1969). Using subsample values as typical values. *Journal of the American Statistical Association*, 64:1303–1317.

Hartigan, J. (1971). Error analysis by replaced samples. *Journal of the Royal Statistical Society Series B*, 33:98–110.

Hartigan, J. (1975). Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values. *The Annals of Statistics*, 3:573–580.

Haslett, J. and Raftery, A. E. (1989). Space-time Modelling with Long-memory Dependence: Assessing Ireland's Wind Power Resource (with Discussion). *Applied Statistics*, 38(1):1–50.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Henry, M. and Zaffaroni, P. (2003). The Long Range Dependence Paradigm for Macroeconomics and Finance. In Doukhan, P., Oppenheim, G., and Taqqu, M., editors, *Theory and Applications of Long-Range Dependence*. Birkhaüser.

Higuchi, T. (1988). Approach to an Irregular Time Series on the Basis of Fractal Theory. *Physica D*, 31:277–283.

Hipel, K. W. and McLeod, A. E. (1978). Preservation of the rescaled adjusted range. 2. Simulation studies using Box-Jenkins models. *Water Resources Research*, 14:509–516.

Hoel, P. G., Port, S. C., and Stone, C. J. (1987). *Introduction to Stochastic Processes*. Waveland Press.

Hosking, J. R. M. (1981). Fractional Differencing. *Biometrika*, 68(1):165–176.

Hughes, M. K. and Funkhouser, G. (2003). Frequency-Dependent Climate Signal in Upper and Lower Border Tree Rings in the Mountains of the Great Basin. *Climatic Change*, 59:233–244.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116:770–808.

Hurvich, C. and Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrics*, 76:297–307.

Hyafil, L. and Rivest, R. (1976). Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1):15–17.

Imbrie, J. and Imbrie, K. P. (1979). *Ice Ages: Solving the Mystery*. MacMillan.

Jaeckel, L. (1972). The infinitesimal jackknife. Technical Report MM 72-1215-11, Bell Laboratories.

Jensen, M. J. (1999). Using Wavelets to Obtain a Consistent Ordinary Least Square Estimator of the Long-memory Parameter. *Journal of Forecasting*, 18:17–32.

Jeong, H.-D. J., Lee, J.-S. R., McNickle, D., and Pawlikowski, K. (2007). Comparison of various estimators in simulated FGN. *Simulation and Modelling Practice and Theory*, 15:1173–1191.

Jones, J. M. (1970). A Comparison of Three Models of Brand Choice. *Journal of Marketing Research*, 7:466–473.

Jones, P. D. and Mann, M. E. (2004). Climate Over the Past Millennia. *Reviews of Geophysics*, 42:1–42. RG2002.

Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes*. Academic Press, Second edition.

Keqian, Z. and Butler, C. J. (1998). A Statistical Study of the Relationship between the Solar Cycle Length and Tree-Ring Index Values. Armagh Observatory.

Klemes, V. (1974). The Hurst phenomenon - a puzzle? *Water Resources Research*, 10(4):675–688.

LaMarche, V. C. (1974). Paleoclimatic Inferences from Long Tree-Ring Records. *Science*, 183:1043–1048.

Lobato, I. and Robinson, P. M. (1996). Averaged periodogram estimation of long memory. *Journal of Econometrics*, 73:303–324.

Mandelbrot, B. B. (1975). Limit Theorems of the self-normalized range for weakly and strongly dependant processes. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 31(1):271–285.

Mandelbrot, B. B. and Taqqu, M. S. (1979). Robust R/S analysis of long-run serial correlation. In *Proceedings of the 42nd Session of the International Statistical Institute*, volume 48 of *Bulletin of the International Statistical Institute*, pages 69–104.

Mandelbrot, B. B. and van Ness, J. W. (1968). Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Review*, 10(4):422–437.

Mandelbrot, B. B. and Wallis, J. R. (1969). Global dependence in geophysical records. *Water Resources Research*, 5:321–340.

Manly, B. F. J. and MacKenzie, D. (2000). A Cumulative Sum Type of Method For Environmental Monitoring. *Environmetrics*, 11:151–166.

Mann, M. E., Bradley, R. S., and Hughes, M. K. (1998). Global-scale temperature patterns and climate forcing over the past six centuries. *Nature*, 392:779–787.

Mann, M. E. and Jones, P. (2003). Global Surface Temperatures over the past two millennia. *Geophysical Research Letters*, 30(15):1820.

Maritz, J. and Jarrett, R. (1978). A note on estimating the variance of the sample mean. *Journal of the American Statistical Association*, 73:194–196.

Martens, M. (2002). Measuring and forecasting S&P 500 index-futures volatility using high-frequency data. *Journal of Futures Markets*, 22:497–518.

Martens, M., van Dijk, D., and de Pooter, M. (2004). Modeling and Forecasting S&P 500 Volatility: Long Memory, Structural Breaks, and Nonlinearity. Tinbergen Institute Discussion Paper TI2004-067/4, Erasmus Universeit, Rotterdam and Tinbergen Institute.

McCoy, E. J. and Walden, A. T. (1996). Wavelet Analysis and Synthesis of Stationary Long-Memory Processes. *Journal of Computational and Graphical Statistics*, 5(1):26–56.

Mikosch, T. and Starica, C. (2003). Long-Range Dependence Effects and ARCH. In Doukhan, P., Oppenheim, G., and Taqqu, M., editors, *Theory and Applications of Long-Range Dependence*, pages 439–459. Birkhaüser.

Milhoj, A. (1981). The Test of Fit in Time Series Models. *Biometrika*, 68(1):177–188.

Miller, I. and Miller, M. (1999). *John E. Freund's Mathematical Statistics*. Prentice Hall, 6th edition.

Mills, T. (2007). Time series modelling of two millenia of northern hemispere temperatures: long memory or shifting trends? *Journal of Royal Statistical Society Series A*, 170:83–94.

Moberg, A., Sonechkin, D. M., Holmgren, K., Datsenko, N. M., and Karlen, W. (2005). Highly variable Northern Hemispere temperatures reconstructed from low- and high-resolution proxy data. *Nature*, 433:613–617.

Montgomery, D. B. (1971). Consumer Characteristics Associated with Dealing: An Empirical Example. *Journal of Marketing Research*, 8:118–120.

Moore, J. J., Hughen, K. A., Miller, G. H., and Overpeck, J. T. (2001). Little Ice Age Recorded in Summer Temperatures from Varved Sediments of Donard Lake, Baffin Island, Canada. *Journal of Paleolimnology*, 25:503–517.

Nunes, L., Kuan, C.-M., and Newbold, P. (1995). Spurious Break. *Econometric Theory*, 11:736–749.

Ohanissian, A., Russell, J. R., and Tsay, R. S. (2008). True or Spurious Long-Memory? A New Test. *Journal of Business and Economic Statistics*, 26(2):161–175.

Overland, J. E., Percival, D. B., and Mofjeld, H. O. (2006). Regime shifts and red noise in the North Pacific. *Deep-Sea Research I*, 53:582–588.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.

Palda, K. S. and Blair, L. M. (1970). A Moving Cross-Section Analysis of Demand for Toothpaste. *Journal of Marketing Research*, 7:439–449.

Palma, W. (2007). *Long-Memory Time Series Theory and Methods*. Wiley-Interscience.

Park, J. Y. (2002). An invariance principle for sieve bootstrap in time series. *Econometric Theory*, 18:469–490.

Peng, C. K., Buldyrev, S. V., Simons, M., Stanley, H. E., and Goldberger, A. L. (1994). Mosaic organization of DNA nucleotides. *Physical Review E*, 49:1685–1689.

Percival, D. B. and Walden, A. T. (1993). *Spectral Analysis for Physical Applications: Multitaper and Conventional Techniques*. Cambridge University Press.

Pettitt, A. N. (1980). A Simple Cumulative Sum Type Statistic for the Change-Point Problem with Zero-One Observations. *Biometrika*, 67(1):79–84.

Ploberger, W. and Krämer, W. (1992). The CUSUM Test with OLS Residuals. *Econometrica*, 60(2):271–285.

Proctor, C. J., Baker, A., and Barnes, W. L. (2002). A three thousand year record of North Atlantic climate. *Climate Dynamics*, 19(5-6):449–454.

Prodan, R. (2003). Pontential Pitfalls in Determining Multiple Structural Changes with an Application to Purchasing Power Parity. Workng paper, Department of Economics, University of Houston, Texas, USA.

Quandt, R. E. (1958). The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes. *Journal of the American Statistical Association*, 53(284):873–880.

Quandt, R. E. (1960). Test of the Hypothesis that a Linear Regression Obeys Two Seperate Regimes. *Journal of the American Statistical Association*, 55(290):324–330.

Quinlan, J. (1993). *C4.5 Programs For Machine Learning*. Morgan Kaufman Publishers, San Mateo California.

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raspopov, O., Dergachev, V., Kozyreva, O., and Kolström, T. (2005). Climate response to de Vries solar cycles: evidence of Juniperus turkestanica tree rings in Central Asia. *Memorie della Societa Astronomica Italiana*, 76:760–765.

Rea, W., Reale, M., and Brown, J. (2007a). Long Memory or Structural Breaks in Temperature and Proxy Time Series. In Oxley, L. and Kulasiri, D., editors, *MODSIM 2007 International Congress on Modelling and Simulation*, pages 3010–3016. Modelling and Simulation Society of Australia and New Zealand.

Rea, W., Reale, M., and Brown, J. (2007b). Modeling Long Memory Time Series: the Shihua Cave Speleotherms. In del Castillo, J., Espinal, A., and Puig, P., editors, *Proceedings of the 22nd International Workshop on Statistical Modelling*, pages 130–135. IDESCAT, Barcelona.

Rea, W., Reale, M., Cappelli, C., and Brown, J. (2006). Identification of level shifts in stationary processes. In Newton, J., editor, *Proceedings of the 21st International Workshop on Statistical Modeling*, pages 438–441. IDESCAT, Galway.

Ripley, B. (2005). *tree: Classification and regression trees*. R package version 1.0-19.

Robinson, P. M. (1994). Semiparametric Analysis of Long-Memory Time Series. *The Annals of Statistics*, 22(1):515–539.

Robinson, P. M. (1995). Log-Periodogram Regression of Time Series with Long Range Dependence. *The Annals of Statistics*, 23(3):1048–1072.

Robinson, P. M. (2003). *Time Series with Long Memory*. Oxford University Press.

Ruddiman, W. F. (2003). The Anthropogenic Greenhouse Era Began Thousands of Years Ago. *Climatic Change*, 61:261–293.

Rudnick, D. L. and Davis, R. E. (2003). Red noise and regime shifts. *Deep-Sea Research I*, 50:691–699.

Salzer, M. W. and Kipfmueller, K. F. (2005). Reconstructed Temperature and Precipitation on a Millennial Timescale from Tree-Rings in the Southern Colorado Plateau, U.S.A. *Climatic Change*, 70:465–487.

Scharth, M. and Medeiros, M. C. (2007). Asymmetric Effects and Long Memory in the Volitility of Dow Jones Stocks. working paper.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Seater, J. (1993). World temperature - trends uncertainties and the implications for economic policy. *Journal of Business and Economic Statistics*, 11:265–277.

Sen, P. K. (1982). Invariance Principles for Recursive Residuals. *The Annals of Statistics*, 10(1):307–312.

Shi, P. and Tsai, C. (2002). Regression model selection – a residual likelihood approach. *Journal of the Royal Statistics Society Series B*, 64(2):237–252.

Shindell, D. T., Schmidt, G. A., Mann, M. E., Rind, D., and Waple, A. (2001). Solar Forcing of Regional Climate Change During the Maunder Minimum. *Science*, 294(5549):2149–2153.

Shuchman, A. and Riesz, P. C. (1975). Correlates of Persuability: The Crest Case. *Journal of Marketing Research*, 12(1):7–11.

Sibbertsen, P. (2000). Robust CUSUM-M-Test in the Presence of Long Memory Disturbances. Technical Report 19/2000, Universität Dortmund.

Sibbertsen, P. (2004). Long memory versus structural breaks: An overview. *Statistical Papers*, 45(4):465–515.

Smith, A. D. (2005). Level Shifts and the Illusion of Long Memory in Economic Time Series. *Journal of Business and Economic Statistics*, 23(3):321–335.

Stephenson, D. B., Pavan, V., and Bojariu, R. (2000). Is the North Atlantic Oscillation a Random Walk? *International Journal of Climatology*, 20:1–18.

Stine, S. (1994). Extreme and persistent drought in California and Patagonia during mediaeval times. *Nature*, 369(3):546–549.

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society*, 39(1):44–47.

Su, X., Wang, M., and Fan, J. (2004). Maximum Likelihood Regression Trees. *Journal of Computational and Graphical Statistics*, 13(3):586–598.

Tan, M., Liu, T. S., Hou, J., Qin, X., Zhang, H., and Li, T. (2001). 2650-year Beijing Stalagmite Layer Thickness and Temperature Reconstruction IGBP PAGES/World Data Center for Paleoclimatology Data Contribution Series No. 2003-050.

Taqqu, M., Teverovsky, V., and Willinger, W. (1995). Estimators for long-range dependence: an empirical study. *Fractals*, 3(4):785–798.

Taqqu, M. S. (2003). Fractional Brownian Motion and Long-Range Dependence. In Doukhan, P., Oppenheim, G., and Taqqu, M., editors, *Theory and Applications of Long-Range Dependence*. Birkhaüser.

Taylor, S. J. (2000). Consequences for option pricing of a long memory in volatility. mimeo.

Teverovsky, V. and Taqqu, M. (1999). Testing for Long-Range Dependence in the Presence of Shifting Means or a Slowly Declining Trend, Using a Variance-Type Estimator. *Journal of Time Series Analysis*, 18(3):279–304.

Thomakos, D. and Wang, T. (2003). Realized Volatility in the Futures Market. *Journal of Empirical Finance*, 10:321–353.

Thomson, D. J. (1990). Time series analysis of Holocene climate data. *Philosophical Transactions of the Royal Society of London A*, 330:601–616.

Tsay, R. S. (1987). Non-Gaussian State-Space Modeling of Nonstationary Time Series: Comment: Detecting and Modeling Changes in Time Series. *Journal of the American Statistical Association*, 82(400):1056–1059.

Veitch, D. and Abry, P. (1999). A Wavelet-Based Joint Estimator of the Parameters of Long-Range Dependence. *IEEE Transactions on Information Theory*, 45(3):878–897.

Wallis, J. R. and O'Connell, P. E. (1973). Firm reservior yield – How reliable are historic hydrologic records? *Hydrological Sciences Bulletin*, 18(3):347–365.

Weedon, G. (2003). *Time-Series Analysis and Cyclostratigraphy*. Cambridge.

Whittle, P. (1953). Estimation and information in stationary time series. *Arkiv für Matematik*, 2(5):423–434.

Wichern, D. W. and Jones, R. H. (1977). Assesing the Import of Marketing Disturbances Using Intervention Analysis. *Management Science*, 24(3):329–337.

Willinger, W., Paxson, V., Riedi, R. H., and Taqqu, M. S. (2003). Long-Range Dependence and Data Network Traffic. In Doukhan, P., Oppenheim, G., and Taqqu, M., editors, *Theory and Applications of Long-Range Dependence*. Birkhaüser.

Wright, J. H. (1998). Testing for a Structural Break at Unknown Date with Long-Memory Disturbances. *Journal of Time Series Analysis*, 19(3):369–376.

Wuertz, D. (2005). *fSeries: Financial Software Collection*. R package version 220.10063.

Zehna, P. (1970). *Probability Distributions and Statistics*. Allyn and Bacon.

Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. (2002). strucchange: An R Package for Testing for Structural Change in Linear Regression Models. *Journal of Statistical Software*, 7(2):1–38.

Zhang, G. Y. (2007). The Distributions of Change Points in Long Memory Processes. In Oxley, L. and Kulasiri, D., editors, *MODSIM 2007 International Congress on Modelling and Simulation*, pages 3037–3043. Modelling and Simulation Society of Australia and New Zealand.

Zhang, L., Mykland, P., and Aït-Sahalia, Y. (2005). A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Date. *Journal of American Statistical Association*, 100:1394–1411.